Spatial hearing rendering in wireless microphone systems for binaural hearing aids

THÈSE N^{O} 7221 (2016)

PRÉSENTÉE LE 11 NOVEMBRE 2016

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

LABORATOIRE DE TRAITEMENT DES SIGNAUX 2

PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Gilles André COURTOIS

acceptée sur proposition du jury:

Prof. J.-Ph. Thiran, président du jury
Prof. P. Vandergheynst, Dr H. Lissek, directeurs de thèse
Dr S. Launer, rapporteur
Prof. T. Francart, rapporteur
Dr C. Faller, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2016

A mes parents
A la mémoire de Serge

Si l'on n'apprend pas à échouer,
on échoue à apprendre.
Tal Ben-Shahar, Apprentissage de l'imperfection

Remerciements

Mes premiers remerciements s'adressent à mon superviseur et co-directeur de thèse, Dr Hervé Lissek. Il y a 4 ans, Hervé m'a proposé un sujet de recherche qui correspondait exactement à ce à quoi j'aspirais, à savoir l'utilisation de l'acoustique et du traitement du signal audio pour améliorer la qualité de vie des personnes malentendantes. Merci à toi de m'avoir soutenu tout au long de ce parcours, y compris dans les moments les plus difficiles, de m'avoir laissé l'autonomie et la liberté de mener mon travail, tout en encadrant ma recherche et en me faisant part de tes intuitions et de ton expérience. Je souhaite aussi remercier mes directeurs de thèse successifs, le Professeur Juan Mosig et le Professeur Pierre Vandergheynst. Juan, j'ai toujours été impressionné par l'excellente ambiance qui règne dans ton laboratoire, grâce à ta bienveillance, mais aussi aux diverses traditions et autres activités de groupe. Je profite de cette occasion pour te souhaiter une excellente future retraite. Quant à Pierre, je te suis extrêmement reconnaissant, comme tous mes collègues, d'avoir sauvé le groupe d'acoustique de l'EPFL. Nos collaborations futures ne feront qu'augmenter davantage la qualité de nos recherches communes. Une mention spéciale à nos deux secrétaires, Eulalia et Mercedes, qui, par leur bonne humeur et leur dévouement (y compris de dernière minute !), contribuent en grande partie au bon fonctionnement de ce laboratoire. Je remercie également l'ensemble des membres de mon jury pour l'intérêt qu'ils ont porté à ma thèse, ainsi que pour les discussions et réflexions enrichissantes dont ils m'ont fait part.

Le groupe d'acoustique du LTS2 est composé de personnes formidables que j'ai eu plaisir à côtoyer durant ces 4 années. Je pense, tout d'abord, aux anciens membres du laboratoire. En premier lieu, Patrick, dont l'expertise en traitement du signal a été une aide précieuse dans la réussite de cette thèse. Je garde d'excellents souvenirs des moments passés avec toi dans le même bureau, des innombrables brainstormings se prolongeant parfois jusque dans la soirée, mais aussi des nombreux délires partagés. Merci aussi à Lukas, Xavier, Cédric et Sami qui ont participés de près ou de loin à mes travaux. Quant aux collaborateurs actuels, Etienne, Romain, Baptiste et Hussein, c'est un véritable plaisir de vous avoir comme collègues au quotidien. Un merci tout particulier d'avoir servi de "cobayes" de nombreuses fois pour la mise en place des tests psychoacoustiques. Du côté des "signaleux", je souhaite remercier Benjamin pour son intérêt pour l'audition et l'acoustique, Johann pour son aide dans les méthodes d'optimisation, et les différents membres de la "geek-room" (ainsi que David) que j'ai plusieurs fois sollicités pour des problèmes informatiques.

Remerciements

Du côté du LEMA, je remercie Jean-François Zürcher, Prof. Anja Skrivervik et Tomislav, pour leur soutien dans la thématique électromagnétique de mon travail, ainsi que Michael, Der Zauberlehrling, pour la traduction de l'abstract en allemand. Merci aussi à Baptiste pour les nombreuses parties du jeu du poisson à l'élastique. Et enfin Joana, Jovanche, Santiago, Anton, Mina et tous les autres pour votre sympathique compagnie dans les locaux du LEMA.

Je tiens aussi à remercier chaleureusement les collègues moratois de Phonak. Tout d'abord Yves, qui a été un excellent chef de projet, ouvert, passionné et disponible. L'équipe DSP, et notamment William (en souvenir de nos journées "Matlab Coder"), Xavier (notre intermédiaire Stäfa/Morat) et Tim. En plus haut lieu, François, Marc et Evert pour l'intérêt porté à nos projets et collaborations. Ce fut un honneur de travailler avec vous tous, mais aussi un plaisir non dissimulé de vous retrouver pour les repas festifs du jeudi midi et les barbecues sous le soleil de Morat. Je termine par Azadeh Ettefagh et Diana Herzog pour leur soutien en IP et procédures éthiques. Un grand merci à Philippe Estoppey, audioprothésiste lausannois de renom, pour ton implication dans notre étude clinique. Je reste impressionné par ton professionnalisme et ta connaissance parfaite de tes nombreux patients. J'en profite d'ailleurs pour remercier chacun d'entre eux d'avoir accepté de participer à cette étude.

Du côté de l'EPFL, j'adresse un grand merci à Sofia, qui a coréalisé avec succès une partie des tests subjectifs rapportés dans cette thèse. Une pensée à Marine, Marina et Marion. Marine qui m'a gracieusement fourni plusieurs "cobayes" de son laboratoire. Marina pour les nombreuses discussions scientifiques, ou autre. Et Marion pour cette belle amitié et ces échanges bio-électro-statistico-philosophiques qui ont animé bon nombre de soirées. Courage à toi pour tes études doctorales imminentes ! Merci à Raphaël pour son expertise en "speech processing", les divers appels téléphoniques, et les échanges de données. Enfin, je remercie sincèrement l'atelier de l'EPFL et l'équipe qui le compose.

Je ne saurais conclure ces remerciements, sans adresser mon infinie reconnaissance envers mes proches. Les amis tout d'abord, les inconditionnels d'Alsace et particulièrement Florence et Gilles, Kalu, mon compatriote suisse, mais aussi les merveilleuses personnes que j'ai rencontrées ici, parmi lesquelles Johannie, Catarina, et Diana (merci pour ta relecture de l'anglais !). La famille également, Cécile, Laurent et Stéphanie, et plus que tout, mes parents, qui m'ont toujours laissé libre de choisir mes orientations et de prendre mes décisions. Je vous remercie également pour votre soutien, particulièrement dans les moments difficiles, et je vous dédie cette thèse. Je termine cette longue liste par celle qui transforme mon quotidien en un véritable bonheur. J'ai beau être un scientifique acharné, je sais qu'aucune équation ni aucun théorème ne pourra jamais expliquer l'amour qui nous unit. Merci à toi Florence, pour hier, maintenant, et tout ce qui nous attend.

Lausanne, le 6 octobre 2016

Abstract

In 2015, 360 million people, including 32 million children, were suffering from hearing impairment all over the world. This makes hearing disability a major worldwide issue. In the US, the prevalence of hearing loss increased by 160% over the past generations. However, 72% of the 34 million impaired American persons (11% of the population) still have an untreated hearing loss.

Among the various current solutions alleviating hearing disability, hearing aid is the only non-invasive and the most widespread medical apparatus. Combined with hearing aids, assisting listening devices are a powerful answer to address the degraded speech understanding observed in hearing-impaired subjects, especially in noisy and reverberant environments. Unfortunately, the conventional devices do not accurately render the spatial hearing property of the human auditory system, weakening their benefits.

Spatial hearing is an attribute of the auditory system relying on binaural hearing. With 2 ears, human beings are able to localize sounds in space, to get information about the acoustic surroundings, to feel immersed in environments... Furthermore, it strongly contributes to speech intelligibility. It is hypothesized that recreating an artificial spatial perception through the hearing aids of impaired people might allow for recovering a part of these subjects' hearing performance.

This thesis investigates and supports the aforementioned hypothesis with both technological and clinical approaches. It reveals how certain well-established signal processing methods can be integrated in some assisting listening devices. These techniques are related to sound localization and spatialization. Taking into consideration the technical constraints of current hearing aids, as well as the characteristics of the impaired auditory system, the thesis proposes a novel solution to restore a spatial perception for users of certain types of assisting listening devices. The achieved results demonstrate the feasibility and the possible implementation of such a functionality on conventional systems.

Additionally, this thesis examines the relevance and the efficiency of the proposed spatialization feature towards the enhancement of speech perception. Via a clinical trial involving a large number of patients, the artificial spatial hearing shows to be well appreciated by disabled persons, while improving or preserving their current hearing abilities. This can be considered as a prominent contribution to the current scientific and technological knowledge in the domain of hearing impairment.

Abstract

Key words: spatial hearing, hearing impairment, hearing aids, assisting listening devices, binaural localization, binaural spatialization.

Résumé

En 2015, le nombre de personnes atteintes d'un handicap auditif s'élevait à 360 millions, dont 32 millions d'enfants, dans le monde entier. Cela fait de la surdité un des grands problèmes sanitaires de notre temps. Aux Etats-Unis, l'augmentation des pertes auditives a été de 160% ces dernières décennies. Pourtant, 72% des 34 millions d'américains présentant un trouble auditif (11% de la population) n'ont pas consulté de médecins, ni utilisé de dispositifs médicaux jusqu'à présent.

L'aide auditive, ou audioprothèse, est la solution la plus répandue pour améliorer l'audition des personnes malentendantes. Elle présente l'avantage d'être non invasive, contrairement aux autres technologies existantes, comme l'implant cochléaire par exemple. Des dispositifs d'aide à l'écoute, qui fonctionnent de paire avec les aides auditives, sont également disponibles sur le marché. Ils constituent une aide précieuse pour faciliter la compréhension de la parole, principalement dans les milieux bruyants et/ou réverbérants. Malheureusement, ces systèmes ne reproduisent pas la dimension spatiale du son, ce qui tend à limiter leurs performances.

L'audition binaurale est une propriété fondamentale du système auditif. Grâce à ses deux oreilles, l'être humain est capable de localiser des sons dans l'espace, d'extraire des informations sur le milieu sonore dans lequel il évolue, de se sentir immergé dans son environnement… En outre, l'audition binaurale est un élément clé contribuant à l'intelligibilité de la parole. Il est donc légitime de penser que l'introduction d'une audition spatiale artificielle dans les dispositifs existants puisse améliorer d'avantage leur efficacité.

Cette thèse étudie et soutient cette hypothèse, selon une approche à la fois technologique et clinique. Elle montre comment certaines méthodes de localisation et de spatialisation du son peuvent être intégrées dans les dispositifs d'aide à l'écoute actuels. En tenant compte des contraintes techniques des aides auditives, ainsi que des caractéristiques du système auditif des personnes souffrant de pertes auditives, cette thèse propose une nouvelle solution permettant de recréer une audition spatiale chez les malentendants. Les résultats obtenus démontrent la faisabilité de cette fonctionnalité et la possibilité de l'implémenter dans les systèmes existants.

Par ailleurs, cette thèse analyse la pertinence et l'efficacité de la spatialisation binaurale sur la perception de la parole. Une étude clinique impliquant un grand nombre de patients a montré que les personnes malentendantes apprécient cette nouvelle fonctionnalité. De plus, les performances actuelles des systèmes d'aide à l'écoute sont améliorés ou préservées.

Résumé

Ces résultats constituent une contribution scientifique et technologique importante dans le domaine de l'audition et des aides auditives.

Mots clefs : Audition spatiale, surdité, handicap auditif, aides auditives, systèmes d'aide à l'écoute, localisation binaurale, spatialisation binaurale.

Zusammenfassung

2015 litten weltweit 360 Millionen Menschen an Schwerhörigkeit, darunter 32 Millionen Kinder. Damit sind Hörbehinderungen ein ernstes weltweites Problem. In den USA nahm während der vergangenen Generationen die Häufigkeit von Hörverlust um 160% zu. Dabei wurden 72% der 34 Millionen betroffenen Amerikaner (11% der Bevölkerung) immer noch nicht wegen ihrer Schwerhörigkeit behandelt.

Es gibt verschiedene Möglichkeiten, Hörbehinderungen zu verringern. Dazu gehören Hörgeräte, die das einzige nicht-invasive und am weitesten verbreitete medizinische Gerät sind. Hörgeräte zusammen mit unterstützenden Horchvorrichtungen stellen ein wirksames System dar, um das verminderte Sprachverständnis bei hörbehinderten Personen zu behandeln. Dies trifft vor allem auf laute und (wider-)hallende Umgebungen zu. Leider geben gewöhnliche Hörgeräte das räumliche Hören des menschlichen Gehörs nur unzulänglich wider, was ihre eigentlichen Vorteile mindert.

Räumliches Hören ist eine Eigenschaft des menschlichen Hörsystems, das auf binauralem Hören basiert. Beide Ohren zusammen erlauben es Personen, Geräusche im Raum zu lokalisieren, Informationen über die akkustische Umgebung zu erfassen, in eine Umgebung ëinzutauchen". Außerdem trägt es enorm zum Sprachverständnis bei. Allgemein wird angenommen, daß die Wiederherstellung einer künstlichen räumlichen Wahrnehmung durch Hörgeräte zur Wiedererlangung eines Teils der Hörleistung einer Person beiträgt.

Diese Dissertation untersucht und unterstützt die zuvor gemachte Annahme durch technologische und klinische Ansätze. Es wird gezeigt, wie bestimmte etablierte Signalverarbeitungsmethoden in existierende Horchvorrichtungen integriert werden können. Diese Techniken stehen in Verbindung mit Geräuschlokalisierung und Verräumlichung. Unter Berücksichtigung der technischen Rahmenbedingungen von modernen Hörgeräten und den Besonderheiten von beeinträchtigten Hörorganen, stellt diese Doktorarbeit eine neue Lösung vor, um für Träger von Hörgeräten die räumliche Wahrnehmung wieder herzustellen. Die erreichten Ergebnisse demonstrieren die Machbarkeit und die mögliche Umsetzung solcher Funktionen in üblichen Hörgeräten.

Außerdem untersucht diese Dissertation die Relevanz und Effizienz der vorgeschlagenen Verräumlichungseigenschaft für die Erweiterung der Sprachwahrnehmung. Klinische Studien an einer großen Zahl von hörbehinderten Patienten zeigen, daß das künstliche räumliche Hören von den Probanden angenommen wird und gleichzeitig ihre Hörfähigkeiten verbessert

oder erhöht. Dies kann als eine wichtiger Beitrag zum aktuellen wissenschaftlichen und technologischen Wissensstand im Bereich der Hörbehinderungen angesehen werden.

Stichworte: Räumliches Hören, Hörbehinderungen, Hörgeräte, assistierende Abhörgeräte, binaurale Lokalisierung, binaurale Verräumlichung

Contents

List of Figures

List of Tables

Concepts & Acronyms

Here is provided the list of the important concepts and acronyms used in this thesis:

Symbols

Symbols	Meanings	Units
α	Type-I error	-
β	Type-II error	-
γ	Integration time (ILD block)	s
δ	Interaural Time Difference	μs
ϵ	Error between the observed and theoretical IPDs	-
θ	Azimuth angle	$^\circ$
κ	VAD threshold	%
λ	Leaky integrator coefficient (RSSID block)	-
ξ	Acceptance threshold (IPD block)	-
ρ	Frame accumulation (Localization block)	-
ϕ	Interaural Phase Difference	rad
$\mathbf{\Omega}$	Probability network (Tracking block)	-
A	Accuracy	%
c	Sound celerity	m/s
D	Euclidian distance	%
f	Frequency	Hz
$\mathfrak{F}$	Fourier transform	-
$\mathfrak{F}^{-1}$	Inverse Fourier transform	-
h	Head-Related Impulse Response	-
H	Head-Related Transfer Function	-
$\overline{H}$	Limited version of H	-
k	Frame index	-
N	Number of samples in a time-frame	-
R_s	Reaction time	s
R	Reactivity	%
s_L (S_L)	Left ear or microphone signal (Fourier Transform)	-
s_R (S_R)	Right ear or microphone signal (Fourier Transform)	-
s_X (S_X)	Body-worn microphone signal (Fourier Transform)	-
$\overline{x}$	Mean value of x	-
$\tilde{x}$	Estimated value of x	-
$\widehat{x}$	Interpolated value of x	-

Introduction

Context

In 2015, the World Health Organization (WHO) reported that 360 million people were suffering from hearing impairment all over the world[1] [33]. This number includes 32 million of children, and represents around 5% of the world population. In the US, one estimates that 1 person over 6 is disabled [201]. In Europe, the proportion of subjects that present a hearing loss (HRL) is about 30% for male adults older than 70 (20% for women), and it reaches 55% for male adults older than 80 (45% for women) [164]. The prevalence of hearing impairment is constantly growing, due to the rise of the world demography, aging of the populations, and development of the industry and noisy leisures [208]. The WHO indeed indicates than 1 over 2 HRL could be avoided if prevention was more expanded [33]. At the time of writing, John Huch *et al.* [100] report that the rate of American people presenting HRL has increased by 160% over the past generations. In particular, the proportion of HI teenagers is 30% higher than it was in the nineties, as a consequence from the increased use of earbuds or headphones.

Hearing disability has lots of consequences on the life of hearing-impaired (HI) people. It mainly affects the communication skills, especially in children, for whom the development of the spoken language is delayed [33]. Moreover, hearing impairment can lead to feelings of loneliness, isolation, frustration, as well as anxiety, and even depression [202]. It is also noticeable when looking at education and employment statistics: HI persons have a poor access to universities, show some significantly greater rates of unemployment, and hold positions with lower grades than normal-hearing (NH) subjects.

Nowadays, there is no medical treatment that can heal a HRL. The major existing solutions consist in the use of external devices. There are 4 main categories of hearing instruments [200]:

1. The middle ear implant. It is made of 2 components: an external device that captures and processes the acoustic signal, and an internal device that simulates the chain of the ossicles in the middle ear,

[1] This concerns adults with hearing loss over than 40 dB at the better ear, and higher than 30 dB for children. Usually, one speaks about a hearing loss when the hearing thresholds are above 20 dB HL. Considering this criterion, the aforementioned proportion of hearing-impaired persons is underestimated.

2. The bone anchored hearing aid, which optimizes and amplifies the bone conduction of the sound up to the inner ear,

3. The cochlear implant, which is the combination of 2 different units. The first is worn behind the ear. It is composed of a microphone and a microprocessor. The second is made of several electrodes that are inserted in the cochlea and that directly stimulate the auditory nerve. The communication between both parts is performed through a wireless connection through the skull. This solution is addressed to profound HI and deaf subjects,

4. The hearing aid (HA), which is a single device that primarily aims to amplify, filter, and output a sound signal in the ear of subjects [135]. HAs are the most widespread solution to improve the hearing performance of HI people.

Hearing aids

The research reported in this thesis is related to HAs. Hearing solutions for disabled persons appeared during the 17^{th} century. From then on, they were continuously improved, following the technological progress. The history of HAs can be split into 6 major ages [58, Chap. 1]:

1. The acoustic age (around 1650) with the use of horn-like apparatus that aimed at collecting the maximum of sound energy to transmit it to the ear of subjects,

2. The carbon age (from 1900 up to the forties) that denotes the system composed of a carbon microphone, a battery and a magnetic loudspeaker (receiver), as shown on Figure 1. This solution brought gains from 20 to 30 dB to the sound signal,

3. The vacuum-tube age (started during the thirties), which saw the rise of one-piece HA reaching acoustic outputs up to 130 dB,

4. The transistor age, resulting from the apparition of the transistor (1952). In 1953, transistor-based hearing devices completely replaced the previous vacuum-tube HA. It made possible to design some adjustable HAs, with basic filtering and dynamic-limiting features,

5. The digital age (started in 1996), that provided advanced signal processing algorithms, easier and precise programming, and size reduction. In 1990, 100% of the HAs were analog [124]. 20 years after, almost 100% of the devices were digital,

6. The wireless age, which is the current period. The communication with other electronic devices (external microphones, TV...) via wireless transmissions (FM and Bluetooth), as well as the development of interconnected HAs, significantly improve the speech intelligibility of HI subjects. In 2014, 82% of audiologists featured wireless technology all over the world, while they were only 74.5% in 2013 [221]. This illustrates the rapid spread of this new HA generation. Furthermore, the aided persons reporting satisfaction with their hearing devices mostly owned wireless HAs in 2015 [2].

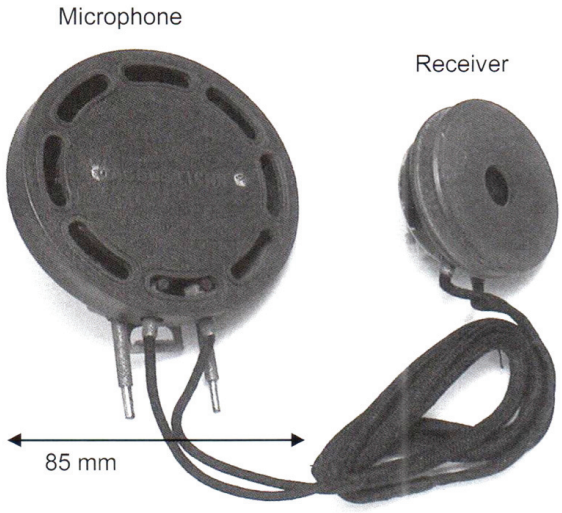


Figure 1 – An example of a carbon HA manufactured in 1902. From [58, page 16].

The benefit from HAs has clearly grew over the past decades. For instance, in 1978, Plomp [190] regretted that HI people took no advantage of a single linear amplification in terms of speech understanding, whereas current HAs, with other amplification approaches, are known to significantly improve the quality of life (speech perception, communications, social integration, self confidence, language and learning abilities in children...) [2, 124, 125, 202, 203]. As an example, 88% of the HA owners reported an enhancement of their life quality thanks to HAs (48% regularly, 40% occasionally) in 2015 [2]. Also, the integration of subjects suffering from HRL is constantly increasing, especially in companies and schools [33].

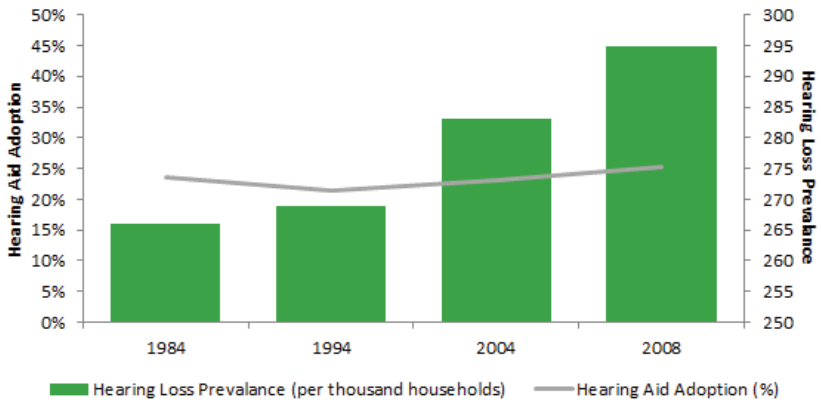


Figure 2 – Evolution of the worldwide HRL prevalence (in million of subjects) and the acquisition of HAs between 1984 and 2008. From [65].

However, the access to HAs is still low. In fact, the worldwide HA adoption rate stagnated at 25% between 1984 and 2008, as revealed on Figure 2, while the prevalence of HRL constantly increases [65]. In the US in 2015, only 20% of the HI persons had ever used HAs. The WHO

highlights that the current production of hearing devices meets less than 10% of the worldwide needs [33], while 1 elderly (65+) out of 3 could benefit from this technology. 11% of the American questioned people perceived hearing difficulties, but only 3.2% of them had HAs [2]. Yet, 10.8 million HAs were sold in 2012 (45% in Europe and 29% in North America), by the so-called "big six" manufacturers covering 98% of the market [228]. They are:

1. Sonova (Phonak, Unitron, Advanced Bionics...), taking 24% of the market share in 2012,
2. William Demant (Oticon, Bernafon, Neurelec...) with 23% of the market share,
3. Sivantos (Siemens, Rexton...) with 17% of the market share,
4. GN Store Nord (ReSound) with 16% of the market share,
5. Starkey (Audibel, Nuear...) with 9% of the market share,
6. Widex with 9% of the market share.

Note that Samsung plans to launch its own HAs in the near future as well [205]. Thus, the hearing-aid industry show to be a powerful and booming business, evidenced by the increase of the US HA sales of 4.8% in 2014 [221] and 10% in Q2 of 2016 [204], and the fact that new generations of HAs are quite a bit more appreciated than the previous ones (e.g. in the US in 2015, 91% of HI subjects having HAs of less than 1 year were satisfied with their instruments, while they were only 74% with 6 years and more devices). The primary impediments and rejections limiting the acquisition of HAs are the cost[2], some unappropriate fittings, the maintenance, and self-attitudes (esthetics, technology reluctance...) [164].

Motivations

The previous section has shown the great benefit from HAs in the HI community. Additionally to the HA contribution to speech understanding and communications, the novel wireless capabilities fasten the development of so-called *assistive listening devices*, i.e. solutions that work with HAs and complement/supplement their performance. The expansion of those solutions is remarkable: in 2015, the resort to these devices concerned 8% of the aided adult HI persons, and 12 % of the 18-39 y.o. disabled people [201]. Note that they are also widely used by children, especially for supporting their integration in classrooms.

Both HAs and assisting listening devices are designed to restore audibility, enhance speech intelligibility, and bring listening comfort. This is often to the detriment of other auditory abilities. In particular, these solutions do not really take into consideration the binaural property of the human auditory system (AS), i.e. the advantage coming from hearing with 2 ears. Binaural hearing is essential for the proper functioning of the AS. It is at the basis of several auditory properties, among which sound localization. The localization of sound

[2] A HA costed around $651 (averaged cost) in 1989, while this cost rose to $2326 in 2008 [65].

sources is related to safety (e.g. localization of sirens in a noisy street), ability to perceive the environments, realism, immersion... Last but not least, binaural hearing strongly contributes to speech understanding. For instance, it facilitates the identification of a speaker of interest in crowded situations. It also helps a fast access to *lip reading*, an essential technique used by HI subjects. Furthermore, several powerful mechanisms for speech intelligibility in the AS are based on binaural hearing.

HAs could dramatically improve if their processing was based on binaural hearing. This is currently one of the main research topics in the HA industry, and it is considered as an essential way for making HAs more attractive [63, 86, 89, 231]. Several new functionalities are developed to allow HAs to focus on the main interlocutor only. The extracted speech from the acoustic surroundings must then be rendered with the adequate spatial cues in the HAs. The interactions between 2 HAs consists in a single system made of so-called *binaural hearing aids* (BHAs) [63], which denotes a class of HAs that communicate, synchronize and share processing through wireless communication. It is supported by the fact that 72% people wearing HAs had bilateral fittings in 2015 [2], i.e. no additional hardware would be required for them.

This thesis goes toward the direction of some new processing techniques working with binaural HAs, based on a specific type of assistive listening devices. The primary applications are for HI children in classrooms, and HI subjects frequenting public spaces such as restaurants, and attending lecture halls, conferences, meetings... As evidenced above, it is directly related to the current and future research in hearing technologies.

Outline and contributions

The thesis follows the chronological work of the research that has been conducted over the past 4 years. It is organized as follows.

Chapter 1 is a literature-based chapter that provides an review about HAs. It is complementary to Appendices A and B that concern the required background in the science of hearing, for both NH and HI subjects. The concepts required for the thesis reading are defined in this chapter. The goal is to avoid readers needing to browse the state-of-the-art in hundred books and articles in order to understand the developments and discussions throughout the next chapters.

Chapter 2 reports the research concerning the development of a binaural localization algorithm for HAs. It aims to simulate the binaural processing done in the AS. The main contributions are the presentation of a **low-cost and efficient process** that meets the specifications demanded by HAs, as well as **the combination of acoustic and electromagnetic cues** to achieve sound localization. An original method of **tracking** is also introduced, and some additional signal processing features are proposed.

Chapter 3 concerns the **optimization** as well as the **final evaluation** of the previously-reported localization algorithm. This requires the acquisition of a wide database of real-world data. Then, a systematic approach is led to maximize the performance of the localization algorithm, which is assessed at the end of the chapter.

Chapter 4 deals with the rendering of an artificial spatial hearing for aided HI subjects, which is a brand new topic in audiology and acoustics. Lots of methods concerning the design of filters that simulate an artificial spatial hearing are reviewed and discussed for the purpose of this research. The primary scientific contribution is the introduction of **the concept of dynamic limitation of spatial filters** that was assessed with a **psychoacoustic study on 40 subjects**. The preliminary subjective evaluation of the spatial rendering via HAs is finally reported.

Chapter 5 is devoted to the core contribution of this thesis, which is the guidance of **a clinical trial on 40 NH and HI patients**, so as to **assess the perception and potential benefit from binaural spatialization methods** on hearing-disabled persons. The spatial rendering provided by the new functionality is assessed in terms of speech intelligibility, sound localization, and preference ratings. The results reveal that the application of spatialization for HI subjects is a promising field of research.

1 Hearing aids

This first chapter presents a review of the state-of-the-art about HAs. It introduces the types, signal processing features, and fittings of such HAs (section 1.1), as well as the impact on sound localization and speech intelligibility (section 1.2), and finally addresses the concept of BHAs (section 1.3) and assistive listening devices (section 1.4). After having gone through this background, the objectives and motivations of the thesis are finally detailed. The references to Appendices A and B should be considered by readers with limited background in hearing science and audiology.

1.1 Functionalities and features

This section describes the different existing types of HAs, the embedded signal processing algorithms available in current devices, as well as the acoustic properties related to earmolds, vents and tubes.

1.1.1 Types of hearing aids

A HA is the most common device that can lessen the effects of sensorineural hearing disorders. It covers mild to profound HRL degrees. Despite the continuous improvement of their performance, it must be clarified from the beginning that current HAs are not able to restore normal hearing [171, Chap. 9].

Over the past 70 years, several types of HAs have been developed. It is primarily the degree of HRL and the discretion of the apparatus that determine the most appropriated kind of HAs for a patient. As depicted on Figure 1.1, 5 primary types of HAs exist:

- *Behind-the-ear* (BTE) HAs (Figure 1.1A), which are dedicated to mild to profound HRLs. In 2012, 20% of the HAs delivered in the world were BTE models [149, Chap. 14].

- *Receiver-in-the-canal* (RIC) HAs (Figure 1.1B), the design of which is closed to that

Figure 1.1 – The different types of existing HAs: BTE (A), RIC (B), ITE (C), ITC (D) and CIC (E). All pictures from www.phonakpro.com.

of BTE HAs, except that the *receiver*[1] is directly located in the ear canal. They have several advantages over BTE models. The receiver is closer to the eardrum, which allows to introduce a smaller gain. The distance between the microphones of the HAs and the receiver is larger. That is good for reducing the risk of mechanical feedback. It comes with a smaller case as well, i.e. more discretion than BTE HAs. Also, no sound is transmitted through the connection tube, which is great to avoid resonances, as discussed later. As a consequence, a finer tube is required, and also contributes to discretion.

- *In-the-ear* (ITE) HAs (Figure 1.1C), which fill the entire concha and the external auditory canal. They are indicated for mild to severe HRLs.

- *In-the-canal* (ITC) HAs (Figure 1.1D), which entirely fill the ear canal. They are recommended for mild to severe HRLs. The major advantages of such HAs are their discretion and the fact that they take advantage from the pinna effects, which is beneficial to sound localization, as discussed in part A.2.1.

- *Completely-in-the-canal* (CIC) HAs (Figure 1.1E), which are inserted deeply in the ear canal, close to the eardrum. The direct consequence is that the amplification can be reduced down to a significant level compared to the other HA models. A second advantage results from the fact that the pinna and concha effects are completely rendered. CIC HAs cover mild to moderate HRLs. HI subjects usually enjoy such kinds of HAs for their total invisibility, despite a more constraining upkeep.

Note that this thesis mostly deals with BTE HAs. Except when notified, the mention of hearing aids only relates to such models.

[1] In the context of HAs, the loudspeaker providing the sound to the ear is called the receiver. It must not be confounded with the RF receiver in WMS.

1.1.2 Signal processing features

The rise of digital HAs over the last decade has led to a great development of new signal processing algorithms that were not feasible with analog HAs. An important number of features are now available. Their presence or absence in the HAs is the rationale for the existence of different ranges of HAs. However, some standard algorithms are included in all devices, whether entry-level or top-end models. Table 1.1 gives an overview of those signal processing features and the issues they are related to.

Signal processing features	**Related issues**
Amplification & Compression	Audibility loss
	Recruitment
	Reduced temporal resolution
Directional microphones	Recruitment
	Reduced frequency resolution
Feedback reduction	HA-induced uncomfort
Noise reduction	HA-induced uncomfort
Frequency compression	Audibility loss
	Reduced frequency resolution

Table 1.1 – Signal processing features available in current HAs and the issues that they have to alleviate. Inspired by [126].

Compression

The amplification is the principal contribution of a HA. This amplification is frequency-dependent and almost always non-linear. This is to face the recruitment phenomenon reported in Appendix B.1.2, and avoid the delivery of an excessive sound pressure level (SPL). Indeed, that would be unpleasant and potentially hazardous for the users. Figure 1.2B depicts the perceptual consequences of an amplification that would be linear. Even if the soft sounds are made audible, the loudness (see Appendix A.1.2 for definitions) of intense stimuli would be unbearable. Compression is applied so that the dynamics of the HA matches the range of comfortable levels for the HI listener. The most used compression scheme is called the *wide-dynamic range compression* (WDRC). There are 2 main reasons for choosing WDRC. First, it reduces the inter-phoneme intensity difference, so that the temporal masking (see Appendix A.1.4 for definitions) is reduced. Second, it avoids reaching the discomfort levels [58, Chap. 6]. To this purpose, it is preferred to peak clipping, which induces noticeable distortions. Multichannel WDRC denotes the fact that compression is performed in sub-bands. The goal is to optimize the compression parameters (compression ratios, thresholds, attack and release times) in a frequency-dependent manner, depending on the HI subject's audiogram and preference.

Directivity

Most current HAs incorporate 2 microphones. Each of them delivers a signal that can be

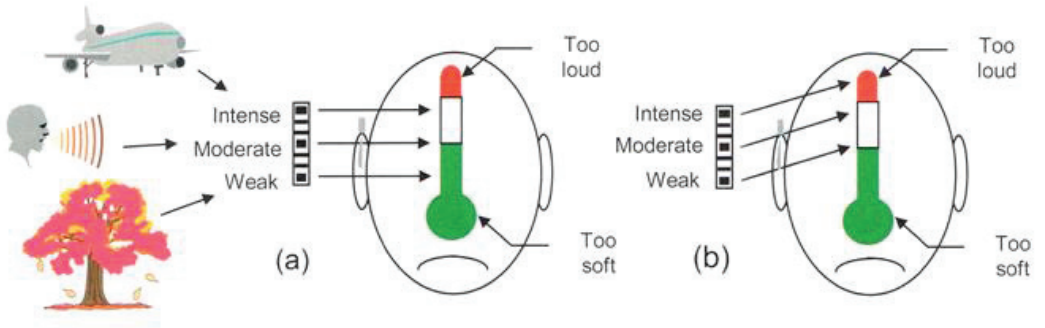


Figure 1.2 – Recruitment phenomenon (A) and the consequences of a linear amplification (B). Adapted from [58, page 3].

independently manipulated in order to provide *directivity*. A first-order subtractive technique is generally used, for which the principle lies in a subtraction of a delayed version of the signal at the second microphone with the signal of the first microphone [58, Chap. 7]. The delay is the combination of the time of flight and a digital added lag. Different values of delay allow to cover the principal directivity patterns: cardioid, super-cardioid and hyper-cardioid. All those patterns decrease the global sensitivity of the microphones, but this reduction is minimized in the front, which increases the signal-to-noise ratio (SNR) in noisy conditions. Indeed, such a processing is based on the assumption that the source of interest is located in front of the aided subject [171, Chap. 9]. The head shadow effect and the distance between the 2 microphones bring about a frequency dependence of the directivity pattern, of which the selectivity increases with the frequency. Adaptive directional microphones are related to the algorithms in which the digital delay is varied in real time, until the maximum of the rear attenuation is reached. This processing is usually performed in a multichannel way.

Feedback cancellation

Martin and Summers [150] define *feedback* as the "*application, to the input of system, of a signal derived from the output of the system. [...] Feedback may lead to instability in a device, leading to uncontrolled oscillation (e.g. the whistle that is often heard from a HA with badly fitting earmold, in which sound from the output is picked up by the microphone)*". One distinguishes 2 types of feedbacks: the *acoustic feedback* (transmitted by sound waves in the air) and the *mechanical feedback* (transmitted by vibrations inside the apparatus). The occurrence of feedback is highly disagreeable and can even be a reason for giving up HAs. Nowadays, feedback is well managed, thanks to the combination of an adequate earmold design (discussed later) and some signal processing strategies. A feedback appears as soon as the amplification of the HAs becomes unstable (divergent) and creates oscillations. This condition is encountered when both the gain and the corresponding phase are positive in a certain frequency area. A simple gain reduction makes the feedback disappear, but it decreases the sound audibility, which is not conceivable.

There are 3 major means of reducing, or even cancelling the ringings [58, Chap. 7]. Formerly,

one attempted to inverse the phase at the oscillation frequency by resorting to all-pass filters. However, current digital HAs introduce too much processing delay for this technique to be sufficiently reactive. The new standards to achieve feedback reduction are the frequency control and feedback path cancellation [87]. The principle of the first is to decrease the gain in the channel where the oscillation occurs, until it disappears. One can also design a notch filter at the ringing frequency. The feedback path cancellation rests on adaptive filter processing. The underlying idea is to approximate and inverse the transfer function (TF) of the leakage path, which can efficiently reduce several simultaneous oscillations [58, Chap. 7]. HAs usually resort to a combination of those techniques.

Noise reduction

Noise reduction processing is primarily implemented in HAs to increase the global comfort and diminish the listening effort. The majority of the algorithms requires to detect the segments that convey only the noisy signal in the sound stream. The goal is to get a reference pattern of the noise. Thus, such algorithms are efficient in environments with steady state noise. Their performance dramatically falls when the noise fluctuates, because the noise pattern constantly changes over time. Inspired by the processing done in the AS, the identification of the noisy segments is based on the analysis of the periodic content temporal fine structure (TFS) processing, see Appendix A.2.1), modulation depth and modulation spectrum (envelope processing) [19]. This allows to detect the presence of speech.

Once done, there are 3 major approaches to perform noise reduction [58, Chap. 7]. The first one is to reduce the amplification or modify the compression in the channels with bad estimated SNRs. Note that it is useless if the noise and speech are located in the same frequency areas. Spectral subtraction is the second technique. Here, the goal is to estimate the spectrum of the noise, and remove it from the speech signal, so that only the signal of interest remains. However, this process has shown to produce audible distortions, mainly because it only affects the magnitude of the signal without considering the phase [19]. Finally the last approach is based on adaptive filters, known as Wiener filters. It aims to estimate the original speech signal (without noise) by adjusting the filter, so that it minimizes the error between the estimated and original signals [19]. The gain of the Wiener filter actually depends on the estimated SNR [58, Chap. 7]. A combination of those techniques is commonly implemented in HAs. Whatever the technique, noise reduction is not supposed to decrease the audibility. This must be guaranteed in each type of processing.

Frequency compression

The last common algorithm that is reported here is the *frequency compression*, to be distinguished from the dynamic compression (WDRC) previously considered. It is common for HI listeners to suffer from a huge loss of the *high frequencies* (HFs). Furthermore, they can present dead regions if the inner hair cells have died in certain parts of the basilar membrane (see Appendices A.1.1 and B.1.2 for the definitions of those concepts). In those cases, even an important gain in the HFs is not sufficient to restore audibility. Additionally the bandwidth

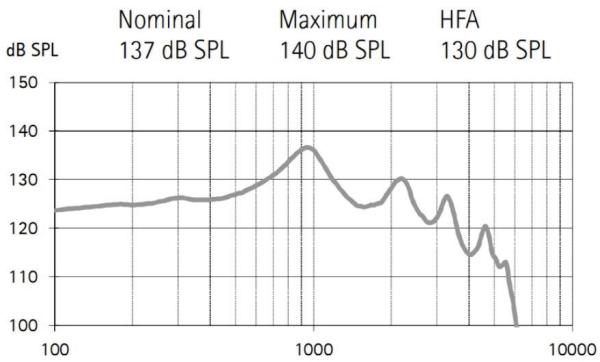


Figure 1.3 – Output SPL of a Phonak Naida Q SP HA, as measured in the 2cc coupler for an input at 90 dB SPL. The device is adjusted to deliver its full gain. From www.phonakpro.com.

of HAs is limited, and the amplification above 6 kHz is usually very low [171, Chap. 9] (see Figure 1.3 showing the output SPL of a Phonak Naida Q SP HA). As the HFs convey prominent information regarding speech intelligibility and sound localization, the resort to frequency compression appears to be the best solution.

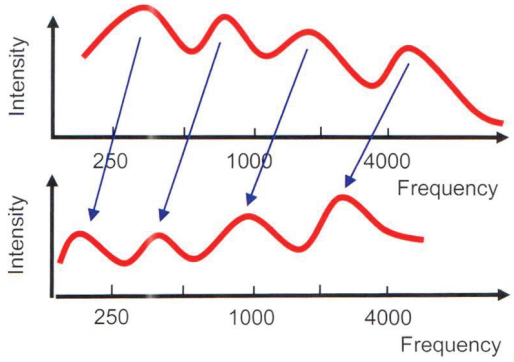


Figure 1.4 – Principle of the linear frequency compression. From [58, page 239].

A linear compression shifts all frequencies to lower frequencies until the HFs become audible again, as shown on Figure 1.4. The primary disadvantage of such a technique is that the shift of the *low frequencies* (LFs) affects the pitch, i.e. a female voice may be perceived as a male speech [58, Chap. 7]. In non-linear frequency compression, the compression arises above a certain frequency threshold. Then, the frequency bandwidth is compressed into a narrower one [166]. This is usually performed via a *Fast-Fourier Transform* (FFT) processing. The principle is to find the bins of maximum energy and shift them in lower frequency ones. Frequency compression obviously requires a certain adaptation time for the users, so that the brain integrates the new frequency map.

1.1.3 Earmolds, vents and tubes

Martin and Summers [150] define an *earmold* of a HA as "*a plug shaped to fit an individual ear, made from an impression of that, through which the acoustic output from a HA is conveyed to the eardrum*". They also indicate that "*the earmold should be seen as an integral part of the HA system as - according to its acoustic design - it can greatly affect the performance of the HA, for better or for worse*". 3 kinds of fittings exist. First, the open fitting, which lets the sound from inside and outside the ear freely flow in and out the ear canal. On the opposite, the occluded fitting completely separates the sound outside the earmold from the internal amplified sound. Between the open and closed fittings, one speaks about vented fitting, i.e. the case where the presence of a *vent* through the earmold allows a part of the sound to move inside and outside the ear canal. The earmold affects 3 main components of a HA fitting, which are: the frequency-dependent gain of the HAs, the probability of feedback occurence, and the self perception of the patient's voice [58, Chap. 5]. That is the reason why the choice of a convenient earmold is prominent to achieve good performance of the HAs, and bring satisfaction to the users.

It is reported in Appendix A.1 that the obstruction of the ear canal, such as the one caused by an earmold, brings about an occlusion effect. The latter is defined as the ratio between the sound pressure from the own voice in the occluded ear and the sound pressure in the open ear [253]. Patients who wear an occluded fitting often complain of a "boomy-like" perception of their own voice. This is particularly reported by subjects whose HRL is greater than 50 dB HL, since they are the primary candidates for a closed earmold. Also, the further the driving in of the earmold, the lesser the occlusion effect. 2 solutions exist to attenuate or cancel the unpleasant occlusion effect. The most common one is to resort to venting. The second one will probably grow in the near future and concerns the digital cancellation of the occlusion effect. It actually requires an additive microphone located in the earmold, which senses the ear canal sound pressure.

When it comes to vents, one is interested in analysing the relationship between the amplified sound path and the vent-transmitted sound path. Dillon [58, Chap. 5] reports that the occlusion effect is perceived even if the vent-transmitted sound path is 10 dB higher than the amplified sound path. The amplified path is affected by the volume present in the ear canal, which represents an acoustic compliance. Such a compliance acts as a high-pass filter. Conversely, the vent path is considered as an acoustic mass that results in a low-pass effect. The combined effect of the compliance and mass thus yields a pass-band shape of the inserted gain (see e.g. [58, page 140]). The peak of the corresponding TF lies in the 1-5 kHz frequency region and may reach around 10 dB. The 2 main advantages of venting is that it helps achieve the target gain, thanks to the resulting gain frequency response, and that it reduces the occlusion effect. It is well admitted that the perceptual reduction of this effect starts with vents of 3-mm diameters [58, Chap. 5]. Note that the presence of a vent enlarges the risk of feedback occurrence. Furthermore, by letting the unprocessed external sound stimulate the tympanic membrane, the benefits from the sound processing in the HA can be significantly

reduced. A trade-off between all these aspects has to be found when dimensioning a vent in an earmold.

Finally, the flexible tube that links the receiver and mold in BTE models brings some effects as well. First, it is responsible for a certain sound leakage, which demands a higher amplification to compensate, and augments the risk of feedback [58, Chap. 5]. Second, it tends to increase the HFs (horn effect). Finally, depending on the length and diameter of the tube, resonances appear and modify the gain of the HA in a frequency-dependent manner. Dampers located at the microphone or at the receiver stage are commonly used to attenuate those resonances.

1.2 Hearing aids: localization and intelligibility

Here are reported the effects of HAs on both speech intelligibility and sound localization. These 2 notions are substantially interrelated, as evidenced in Appendix A.2.

1.2.1 Sound localization

When Byrne *et al.* [29] presented the NAL-NL1 procedure for fitting non-linear HAs, they clearly admitted that "*the frequency response that is optimal for speech intelligibility may not be best for localizing and detecting sounds*". Indeed, the primary objective of HAs is to restore audibility and try to enhance speech intelligibility, rather than improving, or even preserving, an accurate spatial hearing. However, it must be recalled that both mechanisms are substantially related to each other. For instance, Van den Bogaert *et al.* [232] report that there is a slight but significant augmentation of localization errors in the lateral horizontal plane when subjects are aided rather than unaided. The results of this study must be taken with care as only 4 HI subjects have been tested. In a following study, the same authors confirm their result testing 13 HI listeners with BTE and ITE HAs [234]. Conversely, Noble and Byrne [178] observe no difference between aided and unaided conditions for HI subjects equipped with BTE and ITE models.

Bilateral fitting has shown to provide better localization performance than unilateral fittings, considering BTE HAs with occluded earmolds [123]. However, this is untrue for mild HRLs, according to the results reported by Byrne and Noble [28]. As for vents, different sizes are tested and do not change the localization ability of 23 HI listeners [114]. Van den Bogaert *at al.* [234] suspect that occluded earmolds are the main reason to explain the loss of localization accuracy on the sides. They hypothesize that it is due to the absence of access to natural interaural time difference (ITD, see Appendix A.2.1) cues. Tubing, transducers, as well as embedded signal processing add an important amount of delay between the direct sound and amplified sound. This delay can reach 10 ms [58, Chap. 5], whereas the maximal ITD is on the order of 700 μs (see Figure A.10). Moreover, the lag generated by tubing and transducers is frequency-dependent, which brings about the loss of a part of the shape-induced pinna filtering. As soon as the direct sound and the amplified sound merge (i.e. in an open or vented

fitting), some interferences appear and distort the ITD. The delay generated in both HAs can also differ if different signal processing are performed simultaneously in both devices. This happens with unsynchronized and independent HAs, especially when adaptive signal processing is performed [115]. This constitutes another factor of the ITD and interaural level difference (ILD, see Appendix A.2.1) disruption.

As previously discussed, it is expected that the use of BTE, RIC and ITE HAs increases the risk of front/back reversals, since these models bypass the pinna filtering. Conversely, ITC and CIC models would preserve a great deal of monaural cues. Best *et al.* [20] confirm that CIC HAs lead to smaller front/back confusions than BTE HAs. After some accommodation time, the tested subjects actually recover the same performance as when they are unaided. No difference in localization performance between BTE and ITE HAs is noticed in [28, 234]. Dillon [58, Chap. 5] reviews studies that state there is no significant difference in *frontal horizontal plane* (FHP) localization task comparing BTE, ITE and ITC. But BTE effectively yields higher front/back reversals.

When operating independently in both HAs, the WRDC compresses more the signal at the louder ear, while a higher gain is provided to the other HA. Consequently, the ILD range becomes smaller. The more the compression, the more the distortion of the ILD [116]. This phenomenon is emphasized in a multichannel processing, which distorts the ILD and monaural cues depending on which frequency band is processed. However, it appears to have no significant impact on the localization performance of HI listeners, as reported by Keidser *et al.* [113]. The AS seems to adapt to abnormal ILDs and gain mismatches, even though the corresponding ITD/ILD map is modified [28, 113, 245]. Keidser *et al.* [116] observe that this is especially true when broadband stimuli are presented. They hypothesize that it is because the AS generally relies more on the ITD than the ILD in broadband signals. Thus, they state that it is more important to preserve the ITD. Furthermore, Wiggins and Seeber [245] observe that the ILD sensitivity of NH and HI subjects increases with the reduction of the ILD range, i.e. listeners learn to discriminate smaller ILD variations.

Adaptive directional microphone that operates differently in both devices distort the ILD as well [113]. The difference of processing delay required to achieve an optimal directivity is different at both sides, and thus results in an alteration of the ITD [115]. That is the reason why the localization ability of HI subjects is poorer in directional mode than in omnidirectional mode [235]. By applying a frequency-dependent attenuation of the sound coming from the back, directive microphones restore some artificial pinna filtering [58, Chap. 7]. Therefore, the localization performance of HI subjects becomes good again when wearing BTE HAs [113, 115, 171, Chap. 9]. This is at the cost of a reduced accuracy in the lateral horizontal plane. Nevertheless, Keidser *et al.* [114] reports that localization performance is significantly improved after several weeks of adaptation. Again, this evidences the great plasticity of the central auditory system (CAS).

Eventually, the ILD cue is also distorted by noise reduction in independent HAs. This is

because the processing has more effect on the side where the interfering noise is located. It results in an artificial increased ILD, but this has shown to have no influence on localization performance of HI subjects [113].

1.2.2 Speech intelligibility

The primary objective of HAs is to improve speech intelligibility, so that HI subjects recover an easier ability to communicate. As reported in Appendix B.2, the audibility restoration is not sufficient to enhance speech perception, especially in noisy surroundings. Although the localization ability of HI listeners does not seem to suffer from unsynchronized and independent signal processing in both HAs, the distortion of the binaural cues affect speech perception. Some studies reported in [58, Chap. 7] indicate that the head shadow effect and especially the binaural unmasking (see Appendix A.2.2 for definitions) are diminished by the changes occurring with ITD and ILD.

Moore [171, Chap. 9] reviews several studies reporting that non-linear amplification (such as WDRC) improves speech intelligibility in quiet conditions. The reduced temporal masking between vowels and consonants (see part 1.1) is probably the main reason for that. Unfortunately, the beneficial effect of WDRC is lost in complex conditions. Because the compression modifies the speech envelope (i.e. decreases the modulation depth), HAs cannot restore the masking release (see Appendix A.2.2 for definitions), of which the NH listeners benefit from. Additionally, the frequency-dependent nature of WDRC distort the frequency balance of speech signals, and yields a deformation of the phonemes. In particular, the spectrum is flattened and the formant are less distinguishable [58, Chap. 6]. It is particularly true for fast-acting compression that decreases articulation and coarticulation [171, Chap. 9]. Drennan *et al.* [60] presume that the phase distortion operated by multichannel WDRC could be a factor preventing from the enhancement of speech intelligibility. Therefore, they introduce a phase-preserving amplification. However, they fail to demonstrate any benefit for such a processing. This might indicate that the phase changes resulting from WDRC is not an important issue for speech understanding.

Nowadays, directional microphone is the only way to improve the SNR in complex surroundings, especially when the undesired noise is on the sides or behind. It is confirmed by Keidser *et al.* [113], who observe better intelligibility performance in 12 HI patients after several weeks of accommodation. Appleton and König [6] report a 3 dB increase of the speech reception threshold (SRT, see Appendix A.2.2) for 20 HI subjects when directional microphones are switched on, and when diffused noise is rendered. When the noise is presented only on the sides, the SRT increase reaches 5 dB. The major limitations of the directivity appears in reverberant condition, where the interfering reflections come from all directions, or when the speech and noise are located far away from the listener [58, Chap. 7].

The majority of current bilateral algorithms of noise reduction does not provide substantial SNR augmentations, and thus fails to improve intelligibility [171, Chap. 9]. Because it reduces

the amplitude of both speech and noise, the attenuation of the segments with lower SNRs does not succeed in sufficiently increasing the SNR. Nevertheless, noise reduction has shown to improve the listening comfort of HI subjects.

Conversely, non-linear frequency compression has a great impact on speech understanding. Bohnert *et al.* [23] evidence a significant rise of the intelligibility in 7 out of 11 patients. After 2 months, the HI listeners clearly prefer the frequency compression. McCreery [165] observe an immediate better discrimination of consonants in the tested subjects. After 6 weeks, some significant increases in SRTs are reported. In a second study, McCreery *et al.* [166] study the speech recognition in 36 HI subjects, among which 12 are children. Non-linear frequency compression enables to significantly enhance the understanding performance of the listeners.

1.3 Binaural hearing aids

The previous section has evidenced the detrimental effect of the main signal processing features on the binaural cues. The consequence on localization in the horizontal plane has shown to be limited, thanks to the AS adaptation. When it comes to speech intelligibility, it is obvious that a cue-preserving processing could enhance the speech perception. If the preservation of such cues cannot be perfect, it would be expected that the acclimatization to a map of different cues could be easier and faster [193]. Since a couple of years, some great improvements have been achieved from HA manufacturers in this way, with the development of BHAs. Note that all bilateral HAs are not binaural, but all BHAs are bilateral *per se*. The wireless link between both devices is available up to a 30-cm distance, and is around 100 million lower than the US safety limit [193].

The concept of BHAs appeared in the nineties [171, Chap. 9]. The idea was to transmit the signal from one side to the other, no matter the way of doing that. Then, the ITD and ILD would be estimated in sub-bands, using some techniques that are detailed in Chapter 2.2.1. The algorithm would look for the bands where both cues are the smallest. It was hypothesized that such frequency regions would correspond to areas dominated by speech. Conversely, the bands of higher ITD and ILD would be affected by noise and must be attenuated. This would go together with the advantage of directional microphones and would provide a high reduction of signals coming from the sides and rear. Studies reported in [171, Chap. 9] simulate this processing and claim some improvements corresponding to a 5-dB increase of the SNR.

When speaking of BHAs, one must distinguish the *synchronisation-based processing* and the *streaming-based processing*. The first denotes HAs that share single data between each other, such as the volume control level, program selection or compression parameters. The second refers to a new generation of HAs, which exchange full audio streams. The principle of streaming-based HAs is depicted on Figure 1.5. Such BHAs must be considered as a unique system, equipped with 4 microphones and 2 related computational units. From now on, the mention of binaural hearing aids only denotes streaming-based HAs, except when notified. The underlying objective is to take advantage of binaural hearing as the AS does. HAs that

are only synchronized fail to show any benefit concerning speech intelligibility [231], despite a decrease of front/back confusions in a localization test performed by 20 HI subjects [104]. Sockalingam *et al.* [218] evaluate another cue-preserving binaural compression algorithm on 30 HI subjects. They observe a reduction of the localization error in noise. The tested algorithm is also rated by the subjects. They indicate their preference for such a processing in a restaurant-like environment.

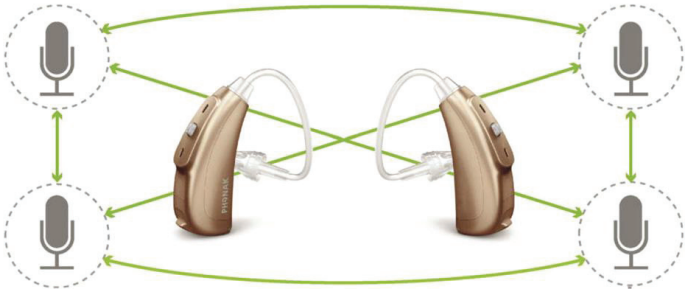


Figure 1.5 – Principle of streaming-based BHAs. From [231].

The availability of a 4-microphone network provides a large room to make directional processing more efficient. Indeed, more advanced beamformer polar pattern can be achieved, providing narrower beams [231]. For instance, the car is an environment where the speech of interest is not on the front but in a predictable location. The directionality can then favour the signals from the sides and can reduce the ones from the front and rear. In the algorithm reported in [130], the system first computes a first order beamformer, as done in a classic HA. Then, the output signal of each device is exchanged between both HAs and a weighted contribution from each of them is rendered. The resulting signal is equivalent to the output of a third order beamformer. Because the distance between the 2 pairs of microphone is artificially augmented, the directional selectivity can be efficient at LFs. This algorithm has shown to improve the SRT of 15 HI subjects by an average value of 1.5 dB compared to a simple first order beamformer. The same kind of algorithms is assessed by Appleton and König [6] on 20 HI listeners. An increase of 1 dB in the SRT is reported over the usual bilateral beamformers. This is similarly observed when the noise is diffused or located only on the sides. Picou *et al.* [187] introduce a beamforming algorithm that preserves the binaural cues. 18 HI subjects have gone through an intelligibility test that shows a large improvement of the speech perception in noise, compared to common directional microphones. However, the efficiency of the algorithm decreases with increasing reverberation.

Yousefian *et al.* [251] developed a noise reduction algorithm that takes advantage of the 4 microphones of the BHAs. They describe it as a simple computational processing that would be easily implemented in a pair of HAs. Their algorithm is based on the calculation of the interaural coherence (IC) between both HAs. A speech signal provides a coherent signal between the 2 ears, while the noise is characterized by a lower IC. They suggest to attenuate the sub-bands where the IC is low (i.e. bad SNR), and amplify the channels in which the IC is high. An intelligibility test over 8 subjects shows that the processing increase the SNR

and results in lower SRTs. This is a great novelty for a noise reduction process. The benefit from the algorithm actually decreases with increasing reverberation, and disappears when a reverberation time of 500 ms is reached. It is regrettable that the tested subjects are all NH listeners, and that the multitalker babble noise comes from a point source in space. This prevents from generalizing the outcome in a context of real-world binaural HA usage by HI people.

1.4 Wireless microphone systems

This last section deals with the so-called assistive listening devices, with a focus on the FM technology.

1.4.1 Principles of existing devices

The previous section has shown that the current algorithms designed for improving speech comprehension are not optimal. This is especially true in reverberant surroundings and in noisy and disturbing areas. In those contexts, HI subjects can rely on the assistive listening devices. They are based on a wireless connection between an emitter and the HAs of a user. The objective is to transmit a speech signal as clean as possible, in order to counteract the adverse effects of noise and reverberation. According to Staab [161], there exist 3 major technologies driving assistive listening devices: infrared, induction and *frequency modulation* (FM). Infrared has become pretty marginal and is not discussed. Nowadays, smartphones can be considered as an additional efficient assistive listening device that offer several applications for encouraging the integration and communication of HI people.

Induction refers to hearing loops. Such loops intend to help HI people understand the speech context in nasty surroundings. The typical use cases are airports, stations, service desks, theaters and churches [112]. This technology requires the presence of a telecoil in the HA. The principle is the following: the voice of a speaker is modulated to generate a large magnetic field in a loop of wires, which induces a current in the HAs of a listener. The speech is then extracted from the electromagnetic signal and rendered in the HAs with a better SNR than it would be if the sound was captured by the microphone of the HAs. Historically, telecoil was developed to capture the magnetic field of telephones so as to get the audio signal. The majority of BTE, RIC and ITE models currently incorporate a telecoil [58, Chap. 3]. This explains why 71% of the users of HAs reported resorting to telecoils in 2014 [112]. Additionally, 86% of them considered the hearing loop as the best assisting listening system, but it is unsure whether they experienced other technologies before. It is noteworthy that NH subjects were shown to enjoy telecoils in the same study. The reasons for this are that hearing loops enhance speech intelligibility, reduce listening effort, and increase sound quality and plesantness. The major advantage of induction over FM is its reduced battery consumption and the low cost. On the other hand, it is very sensitive to parasitic noise (switch of power supplies, 50 or 60-Hz hum...).

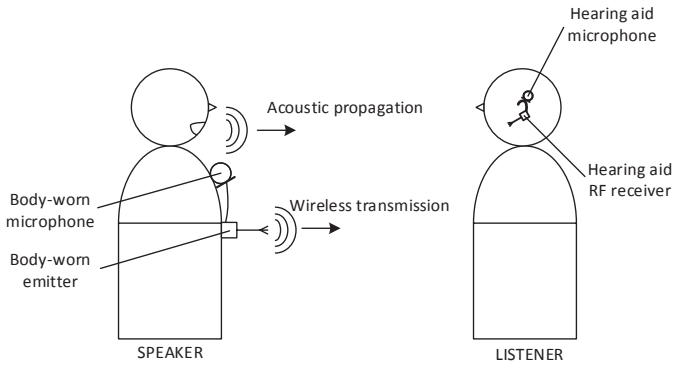


Figure 1.6 – Principle of the FM assisting listening systems. From [50].

FM systems appeared 30 years ago. Currently, a typical system consists of a small transmitter microphone, which picks up the voice of a speaker, and sends the speech signal wirelessly to a RF receiver plugged into the HAs of a listener. The *radio-frequency* (RF) receiver is connected via the *direct audio input* (DAI) of the hearing device. It can also be integrated in the HA. The principle is shown on Figure 1.6. Since the voice is picked up very close to the speaker's mouth, and thanks to a beamforming processing, almost only the direct sound is recorded. The common use cases of such systems include classrooms, lecture halls, auditoriums or restaurants [161]. The objective is to ensure a high-quality reproduction of the sound whatever the distance between the speaker and HI subject. Indeed, with a purely acoustic transmission, the sound intensity diminishes when the distance increases. The consequence is that both the SNR and direct-to-reverberant ratio (DRR) decrease, as shown on Figure 1.7. The FM systems circumvent this problem, guaranteeing a constant SPL whatever the distance (up to a certain critical distance, where the FM power is not enough to guarantee a proper reception of the radio signal, usually above 15 m) [163]. The FM solution is a bit tough to use, especially because of the proprietary wireless technologies that brought incompatibilities and issues for users to select the good frequency bandwidth [162, 163].

The FM is based on a modulation of the speech on a carrier frequency. The demodulation takes place in the RF receiver. Different carrier frequencies enable to avoid interferences between different emitters, and enable the possibility for several talkers to wear an emitter microphone. The signal conveyed via FM can be combined with the local HA microphone signal (FM+M mode) or not (FM-only mode). HI subjects, and especially children largely preferred FM+M is rather than FM-only, because the latter provokes a feeling of detachment from the environment. However, the mixing of the FM with the acoustic signal reduces the benefit from the FM signal. The FM can improve the SNR up to 20 dB [58, Chap. 3]. Therefore, the level of FM-transmitted signal is more amplified that the M-transmitted signal. This is called the *FM advantage*, expressed in dB. Values from 10 to 20 dB are usual. In *dynamic FM*, the FM advantage is automatically adjusted depending on the quality of the sound captured

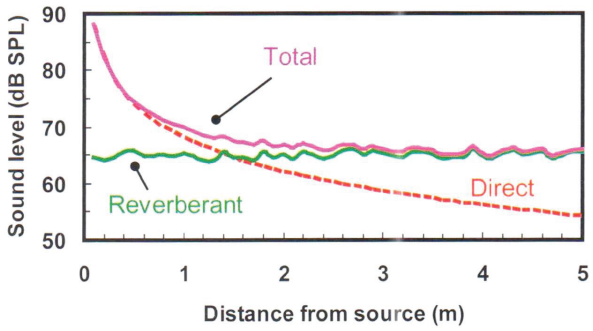


Figure 1.7 – Variation of the SPL as a function of the speaker-to-listener distance, for the direct sound (red), the reverberated sound (green), and the combination of the 2 (pink). From [58, page 56].

by the microphone. Note that the activation of the FM yields the deactivation of one of the 2 HA microphones, because only 2 inputs are available in current HAs. Hence, no directivity is possible but future models will overcome this issue.

The denomination “FM systems” is actually obsolete. In fact, all new assistive listening devices rest upon a *digital modulation* (DM), while the FM transmission tends to disappear. The DM (Roger) technology is based on 2 techniques, which are the frequency-shift keying and the frequency-hoping spread spectrum. The principle of the first is to provide small variation of the frequency around 2.4 GHz in order to code a value of 1, whereas no frequency change occurs for a 0 value. Then, the frequency-hoping spread spectrum denote the fact that both emitter and receiver hop to different carriers between 2.4 and 2.482 GHz at a certain rate. This is to avoid interferences with other emitting device. The DM operates at a frequency of 2.4 GHz, i.e. the same frequency area as the bluetooth protocol (2.4-2.48 GHz) and Wi-Fi 802.11b and 802.11g. The expression of *wireless microphone systems* (WMS) is adopted throughout the thesis to denote both FM and DM transmissions.

1.4.2 Speech intelligibility and speaker localization

It has been a long time since WMS has shown to provide significant and prominent advantages for speech intelligibly in HI users. The first studies reporting this enhancement were published at the beginning of the eighties, see e.g. [95]. WMS rely on the *binaural summation* of a *diotic* signal (i.e. exact same signal at the 2 ears) [58, Chap. 15]. The signal is then processed as if it was picked up by the HA microphones. Crandell and Smaldino [53] report a SRT reduction of 10 to 20 dB thanks to the FM technology. Another study [68] is interested in the understanding difference between the FM-only and FM+M modes. The FM-only rendering significantly oversteps the performance obtained with the FM+M mode. Lewis *et al.* compare the speech intelligibility in 44 HI subjects between the usual directional processing of the listeners and the FM-only mode. They observe a significant reduction of 19 dB of the average SRT in the

second case. Nevertheless, they indicate that the outcomes may not be the same for children.

Recently, Thibodeau [229] has evaluated the effect of dynamic FM on speech perception. She mentions some FM advantage values up to 24 dB. The speech signal was delivered at 84 dB (A). As soon as the noise level is upper than 63 dB (A) (i.e. a SNR of 21 dB), she reports a maximum improvement of 50% of the speech recognition score (SRS, see Appendix A.2.2) averaged on 10 HI listeners over the static FM. In very noisy condition (SNR of 4 dB), dynamic FM outperforms the usual FM by 22.7% of SRS. Regarding the results they have obtained for a panel of 11 children suffering from autism spectrum disorders or attention-deficit hyperactivity disorders, Schafer *et al.* suggest that FM systems allow impaired children to reach the same understating performance than a control group of 11 normal children. Also, the classroom behavior rated by the teachers improves. It shows that WMS may not be only useful for HI subjects.

When it comes to localization, it is obvious that all the binaural and monaural cues are removed in WMS, when use in the FM-only mode. Indeed, the body effects are bypassed, and the diotic rendering prevent the reproduction of the binaural cues. This issue is at the root of the isolation complaint previously reported. The FM+M mode may bring back a sense of immersion, but this is at the cost of a reduced intelligibility. The next section is concerned with the objective to recover an accurate spatial hearing in the context of WMS. That is the core of this thesis.

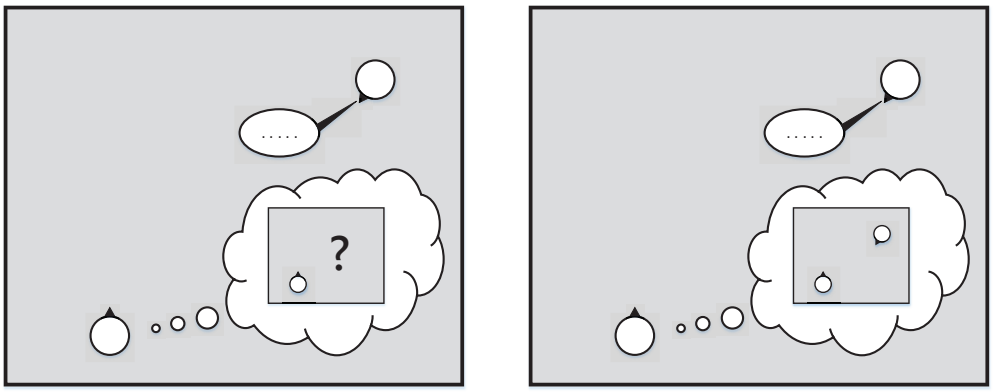
1.4.3 Improvement of current systems for speaker localization

Concept

The previous section has emphasized the great benefits from WMS on speech understanding, as well as the resultant suppression of the spatial hearing. In Appendix A.2, the contribution from speaker localization for lip reading is reported. Let imagine a situation in which a HI subject attends a conference. There are e.g. 5 speakers in front of the listener, all wearing a body-worn microphone. The voice corresponding to the current talker is rendered in the HAs, then another person is speaking, and the respective voice is delivered as well. All signals are clean, with a high SNR and low reverberation. In his daily life, the subject always uses the combination of sound and lip reading, since the acoustic signal is not enough to understand the speech content. Let also recall that the discrimination of pitch is altered in his AS (see Appendix B.2.2). When the new speaker takes the floor, he has to find which of the 5 speakers is currently talking, and cannot relies on a precise pitch discrimination. This represents a loss of time that prevents him for understanding the beginning of the talk. Even worse, let consider the case where all speakers exchange one after the other in a short time. It becomes totally impossible for the HI listener to follow the conversation. What is missing? The clues coming from the acoustic signal in his HAs to infer the location of the different speakers.

Another interesting scenario is the situation where a HI pupil performs a dictation in a classroom. The child exploits lip reading as well, in order to catch on the words he does not understand with his ears. Of course, he cannot look at the teacher all the time because he

has to write in his copybook. While reciting the dictation, the teacher is moving across the classroom. In the HAs, the sound is static and does not reproduce the motion of the speaker. That is, each time the pupil raises the sights, the position of the teacher has changed and the child needs time to find the speaker again, as shown on Figure 1.8A. A time lost to read on the teacher lip.



(a) (b)

Figure 1.8 – A typical use case of a current WMS (A). The targeted solution of the thesis (B). From [51].

The objective of this thesis is to suggest a solution that solves that issue. At the end of the research, the HI users of a wireless microphone system would be ideally able to acoustically perceive the location of the single or multiple emitter(s) worn by the different speakers, as depicted on Figure 1.8B. 2 recent patents (2016) from Oticon [109] and Starkey [110], goes toward the same direction with similar applications. This shows that the current research is trendy and highly competitive. Besides, Farmani *et al.* [71] have just published a scientific paper based on a part of the work presented in this thesis.

Practically, the thesis has to propose a way to localize the speaker(s) in the FHP relative to the listener (Chapter 2). The localization performance is optimized and assessed with real-world data (Chapter 3). This is the first original contribution of the thesis, i.e. the developement of a localization algorithm working in real time and in adverse environments, compatible with HA specifications. Then, the voice of the current talker captured by the microphone of the emitter must be rendered as if it is coming from the speaker's determined location (Chapter 4), which introduces another contribution consisting in the development of a spatialization algorithm evaluated on 38 NH listeners. An additional information of the distance between the speaker and listener would be interesting, but not mandatory. Finally, the efficiency of the algorithm has to be assessed on a large panel of HI subjects (Chapter 5). The evaluation of the binaural spatialization technology on HI subjects is actually the main scientific contribution of the thesis.

Real-world constraints

HAs are devices with limited embedded memory and processing power. Thus, the algorithms that require an important amount of stored data cannot be implemented. The computational cost should also be circumscribed, for 2 main reasons. First, too many operations would make the processor fall behind the real-time framework and eventually crash. Second, a simple and fast algorithm would be advantageous in terms of battery life [50]. Practically, the prototype is made of a *body-worn unit* (BWU) that includes an Atmel ARM9 CPU, which requires that the processing power is lower than 3 million instructions per second. The RAM usage must be less than 4 kilobyte. The sampling rate is 16 kHz and the time frame is 2 × 4 ms (128 points). The distinction between the frame rate of the incoming frames (4 ms) and the frame rate of the analysis frames (8 ms) is fundamental to understand the functioning of the localization and spatialization processes. Indeed, a new frame arrives every 4 ms, but 2 consecutive frames are concatenated to form a 128-sample analysis frame. Finally, the developed algorithm is asked to limit at most the exchange of binaural information between both HAs, in order to limit the battery consumption [50, 181].

The targeted acoustic environments are far from anechoic or quiet conditions. Reverberation, especially strong early reflections, may occur and disturb the process. This would lead to significant errors if the localization focuses on the incidence direction of the reflections, instead of considering the direct sound [50]. Interfering noises (air conditioner, beamer rumble, competing speakers, coughs...) are likely to take place as well. Strategies to counteract these issues have to be found.

Available signals

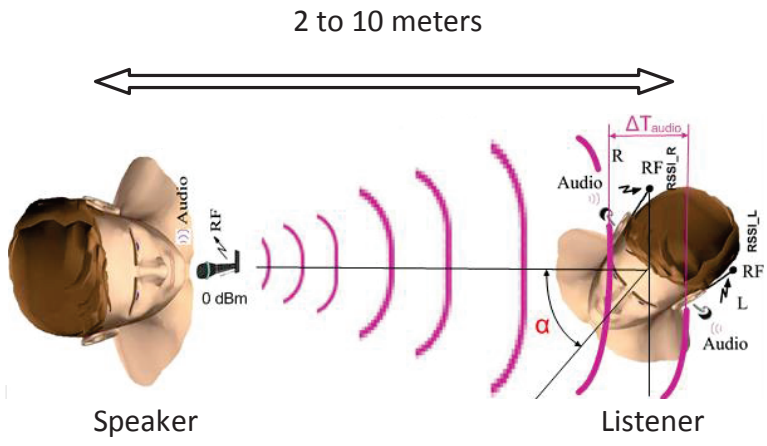


Figure 1.9 – The situation considered in this thesis, with all available signals. Adapted from [70].

The whole system depicted in Figure 1.9 brings access to different acoustic and radio-frequency

(RF) signals:

- The audio signals captured by the HA microphone at both sides. These signals are written s_L and s_R in this thesis. They are degraded by noise and reverberation,
- The demodulated audio signal coming from the microphone of the emitter. It is denoted by s_X. It is the clean speech picked up at the mouth of the speaker. This signal must be spatialized at the end of the process,
- The *received signal strength indication* (RSSI) that represents the strength of the RF signal that reaches the left and right HAs. These are written as RSSI_L and RSSI_R.

All these signals are exploited so as to estimate the location of the speaker, which is the object of Chapter 2.

2 Development of a binaural localization algorithm

This chapter presents the first step of the research, which is related to the localization of the speaker(s) wearing the body-worn microphone. The final goal is to develop a process that is able to localize the speaker and provide their position in real time, using all the available cues. The main contributions of this chapter are the combined use of acoustic and electromagnetic information to achieve the localization, as well as the adaptation of techniques from the literature to the constraints of HAs. As discussed in what follows, the primary difficulty is the impossibility to directly compare the left and right acoustic signals, which demands the resort to smart strategies to be overcome. Even under adverse conditions, the algorithm must deliver the most accurate and stable output as possible, despite the necessity to perform a low-cost signal processing. This goes through some evaluation and decision steps integrated in several stages of the algorithm, as well as the implementation of a simple and efficient tracking procedure.

Part 2.1 introduces the applications and some methods for sound localization reported in the literature. Then, part 2.2 details the computation of the acoustic and electromagnetic cues, of which the combination is the object of part 2.3. Part 2.4 relates to the tracking procedure. The conclusions of this chapter are drawn in part 2.5. Note that the reported algorithm was patented in 2015 [48].

2.1 Introduction

This first part presents the past and present applications of the localization algorithms. Then, the mathematical and computational methods reported in the state-of-the-art related to the *binaural localization algorithms* (BLAs) are described.

2.1.1 Applications

Algorithms performing localization of sound sources have a wide field of applications. The military industry was the first to invest in the development of such algorithms for radar and

sonar technologies [137]. Among the early domains of civil applications of sound localization techniques, robotics appears to be the most prominent one. The aim is to improve the navigation of robots [119] or to facilitate the interaction between humans and robots [248]. In this context, it is common to resort to a minimum number of 3 microphones to detect and track sound sources in space [119]. The number of embedded microphones is usually not a strong constraint. A large number of microphones constitutes a *microphone array*, for which the underlying signal processing strategy is most commonly *beamforming*. It consists in "steering" the beam of the array in different locations, in order to find the direction that brings the maximum amount of energy [75, 137]. The higher the number of sensors, the narrower the beam of the beamformer. This method is used in various fields dealing with speech enhancement (noise reduction) [67], security (video monitoring) [237], uncrewed vehicle (accurate navigation and environment monitoring) [145], computational auditory scene analysis (sound separation) [239] or video conferencing (to focus the camera on the current speaker's face) [99].

The major drawback of the beamforming technique is that it is computationally expensive, because a great number of signals and operations needs to be processed in real time [239]. With the objective to reproduce the excellent performance of the AS as a sound localization system, one attempts to develop an efficient multi-source localization with only 2 microphones [117]. This is the case in humanoid robots, that are at their early stages of development [67, 103]. Willer *et al.* [248] indicate that such kinds of processing have potential applications in BHAs, under some *a priori* hypotheses. The algorithm developed and introduced in this chapter rests on 2 major assumptions. First, only one source is active in the environment (1 speaker or more speakers successively talking). Second, this sound source is located in the FHP.

2.1.2 Methods

Over the past decades, a significant number of BLAs have been reported in the literature. They can be classified into 2 main groups: those that are based on a head-related transfer function (HRTF, see Appendix A.2.1) database, and those that first estimate some spatial cues so as to achieve localization. Both types of algorithms aim to retrieve the *direction of arrival* (DOA) of the sound.

The primary advantage of resorting to HRTFs is that it allows to localize in both azimuth and elevation. The inverse HRTF filtering method [119] attempts to find the inverse HRTF-based filter that must be applied on the left and right received signals in order to equalize them. The "source cancellation algorithm" [117] maximizes the cross-correlation between the quotient of the left and right signal spectra and a set of HRTF ratios in various directions. In the "cross-channel algorithm" [145], the signal recorded at the left ear is filtered with a set of HRTFs from the right ear and conversely, until the left and right filtered signals sufficiently match. In all these algorithms, the pair of HRTFs that satisfies the adequate criterion gives access to the DOA. Talagala *et al.* [226] apply subspace decomposition on the input signals and on the

HRTF database. The orthogonal property of the adequately chosen subspaces is then used to estimate the current location of the sound source.

The algorithms of the second family first compute one or several spatial cues. Then, these observations are either compared with a set of reference values, or fed into a mathematical model or a neural network, and the DOA is inferred. The majority of these algorithms performs localization only in the FHP. The BLA proposed by Lim and Duda [139] derives the ITD and ILD at the output of a cochlear model, and compares them with a collection of ITD and ILD values for different DOAs. Li and Levinson [138] also resort to a cochlear model. They compute the short-time ITD, ILD and spectral cues from the interaural differences and intra-aural variations. These estimated cues are the input of a statistical model, previously trained with experimental data. In the BLA reported by Nix and Hohmann [177], a Bayesian classifier compares the derived ITD and ILD with histograms of these cues collected in various acoustic conditions, and looks for the most likely DOA. Zhou *et al.* [252] extend this algorithm for moving sound sources. A bio-inspired model is introduced by Willert *et al.* [248], based on binaural cues extracted from some cochleagram representations compared with some reference maps of ITD and ILD. Instead of using some reference data, Brandstein [26] exploits a mathematical model to infer the DOA from the observed ITD. As reported in Appendix A.2.1, several formulas have been proposed to match the ITD with a corresponding azimuth in the FHP, e.g. the Woodworth's formula (Equation A.1) or the sine law (Equation A.2). Raspaud *et al.* [199] tune the Woodworth's formula with empirical scaling factors. Eventually, the estimated spatial cues can also serve to feed a neural network [56, 103, 174, 212]. Note that some BLAs combine several techniques resorting to both families, as in the algorithm of Wan and Liang [239], who associate an estimation of the ITD prior to the use of the "cross-channel algorithm" proposed by MacDonald [145].

It is uncertain how the previously described algorithms behave in real acoustic conditions. Indeed, it is remarkable that the assessments of the BLAs are often performed under quiet and anechoic environments. Only few authors report the evaluation of the robustness of their procedure against noise and reverberation, attempting to find solutions to circumvent their effects. The approach of Nix and Hohmann [177] seems to be the strongest one at dealing with the interfering noise. However, it requires a large amount of stored data and the estimation of the SNR, which is not a trivial task.

In the context of this thesis, the algorithms requiring the storage of a HRTF database are unapplicable, due to the lack of embedded memory. Thus, the orientation towards the second families of algorithms is favored, although simpler methods must be found. Also, one has to make sure that the performance of the algorithm is preserved in complex surroundings.

2.2 A multi-cue algorithm

This part introduces each localization cue derived in the BLA. The computations are based on some techniques reported in the literature and the choice of the most adapted ones. Because

the algorithms resort to several cues, which are either acoustic or electromagnetic ones, one speaks about a *multi-modal approach*.

2.2.1 Interaural phase difference

Previous work

In a recent literature review, several methods to extract the ITD were reported [111]. The most famous technique consists in deriving the *interaural cross-correlation* (IACC) and in finding the delay associated with its maximum value. Under free-field conditions, the left and right ear signals s_L and s_R, received from the emitted signal s with a DOA θ, can be modeled as:

$$\begin{cases} s_L(t) = h_L(t,\theta) * s(t-T) \\ s_R(t) = h_R(t,\theta) * s(t-T) \end{cases}, \tag{2.1}$$

where h_L and h_R denote the left and right *head-related impulse responses* (HRIRs) for a sound source at the azimuth θ, T is the acoustic time of flight between the source and head, and $*$ denotes the convolution operator. In the LFs, the effect of the HRIRs can be approximated by a pure time-delay. Therefore, the following relation holds:

$$s_R(t) = s_L(t-\delta), \tag{2.2}$$

where δ denotes the ITD. In this thesis, the ITD is taken positive when the sound source is located on the left (positive azimuths) and negative when it is on the right (negative azimuths). The IACC is defined as follows:

$$r_{LR}(\tau) = \int_{-\infty}^{+\infty} s_L(t) s_R(t-\tau) dt. \tag{2.3}$$

Then, the ITD can be estimated as:

$$\tilde{\delta} = \arg\max_{\tau} r_{LR}(\tau). \tag{2.4}$$

The interaural phase difference (IPD) is estimated by extracting the phase of the *interaural cross-spectrum* (IACS) between S_L and S_R, the Fourier transforms of the signals s_L and s_R. From Equation 2.2, the following relation can be written:

$$S_R(f) = S_L(f) e^{-2\pi f \delta}. \tag{2.5}$$

The cross-spectrum is defined as:

$$R_{LR}(f) = S_L(f)S'_R(f). \tag{2.6}$$

Replacing S_R according to Equation 2.5, Equation 2.6 becomes:

$$R_{LR}(\tau) = ||S_L(f)||^2 e^{2\pi f\delta}. \tag{2.7}$$

Let φ be the phase of the cross-spectrum, i.e. the IPD. The ITD can be recovered from the IPD:

$$\varphi(f) = 2\pi f\delta \Longleftrightarrow \delta = \frac{\varphi(f)}{2\pi f}. \tag{2.8}$$

The estimation of the ITD or IPD is a tough task in real acoustic environments, i.e. when background noise and reverberation are not negligible. In those cases, the previously reported methods fail to produce an accurate and stable output. To counteract the effects of reverberation, Huang *et al.* [99] suggest to compute the IACC only in the areas with high DRR. Vieira and Almeida [237] use the same method, introducing an onset detector in order to extract the time segments where the ITD estimation is likely to be reliable. The major limitation of this technique is the need to account for predetermined parameters for each room.

The *phase transform* (PHAT) [122] is known to significantly improve the performance of the IACC method in the presence of reverberation. This technique tends to increase the peak of the dominant delay in the reverberated signal, reducing the risk of detecting peaks corresponding to reflections. However, the PHAT also highlights the components of the spectrum with poor SNR, which makes it less powerful in noisy surroundings. Branstein [25] suggests to emphasize the weight of frequency components that exhibit the highest periodicity in the IACC. He assumes that the portions of speech signal with a clean periodic nature are less degraded by noise and reverberation. This approach was shown to outperform the PHAT in a noisy and mildly reverberant simulated room. Hwang and Choi [103] apply the IACC method on a sparse representation of speech signals in order to determine the delay between the main spikes in a time-frequency pattern, obtained from a modified matching pursuit algorithm. Recently, there were attempts to use the statistical distribution of speech signals to estimate time delays, such as a generalized Gaussian distribution [186] or a Laplacian mixture model [67]. All these methods have been tested for different acoustic environments and are shown to be robust against noise and reverberation.

Methods based on the IACS and the determination of the IPD have also been improved to manage noisy and/or reverberant surroundings. Based on the assumption that the phase of the IACS is linear (after Equation 2.8), the algorithm of Li and Levinson [137] estimates the

slope of the phase spectrum as the estimation of the ITD. This technique has been validated under simulations of an anechoic environment with SNRs from 40 down to 20 dB. In order to circumvent the effect of reverberation, Fujii *et al.* [75] suggest to derive the cepstrum of the warped phase of the IACS so that only the components showing an oscillation rate lower than 1 ms (i.e. corresponding to the direct sound) are considered. Some measurements have been conducted using speech signals in a quite reverberant room to assess and confirm the efficiency of this approach.

Suggested approach

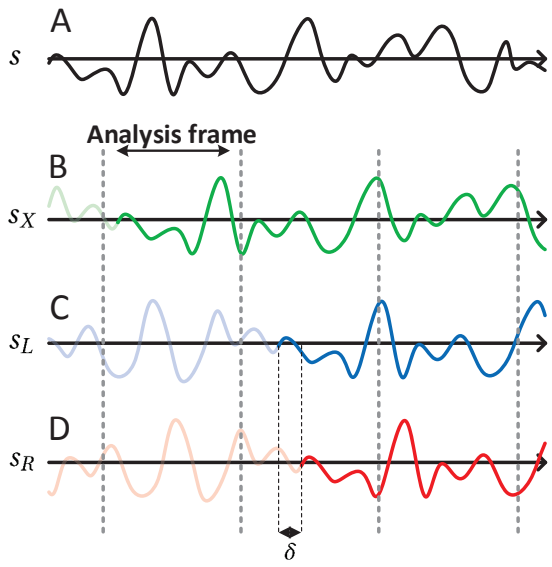


Figure 2.1 – Example of a configuration where the speaker-to-listener distance makes the original speech signal (A) be rendered first via the wireless transmission (B), then via the microphone of the left (C) and right HAs (D). The speaker is on the left relative to the listener. There is no common sample in the same analysis frame between s_X and the audio signals s_L and s_R.

As mentioned in Chapter 1.4.3, the present application prevents a direct comparison of the left and right signals, either to compute the IACC or IACS, because the audio signal of the left HA is not available in the right device, and conversely. This issue can be overcome by using the demodulated audio signal from the radio transmission as a reference, because it is exactly the same in the 2 HAs. The simplest trick is to compute the delay between the demodulated audio signal s_X and the signals captured by the microphones s_L or s_R in both devices. The ITD can then be recovered differentiating these 2 delays [50]. The main advantage of this technique is that it only requires the exchange of a single value between both HAs for each analysis frame.

In the targeted hardware, the length of an analysis frame is 8 ms, corresponding to 128 samples at 16 kHz. The delay for the modulation, transmission and demodulation of the signal s_X

is fixed and close to 20 ms, while the delay due to the acoustic time of flight and processing latency of the audio signals s_L or s_R is between 10 and 36 ms for a speaker-to-listener distance between 1 and 10 m. Therefore, it may happen that the current analysis frames from the demodulated audio and microphone signals do not share any common sample. This situation is depicted on Figure 2.1. The signals s_L are s_R are so late that there is no common pattern with the s_X signal in the same time frame. Figure 2.2 shows the percentage of common samples between the radio and audio frames (the left or right, depending on the considered HA) as a function of the speaker-to-listener distance. The compensation of the radio and audio delay is reached for a distance between 4 and 5 m for a 128-sample frame size. The distance range that provides more than 50% of common samples is approximately from 3 to 6 m. It is assumed that below 50%, the cross-correlation does not cover a sufficient delay range. Only a frame length of 512 points would be enough to cope with a 10-m span. However, the memory usage constraints do not allow to buffer 4 successive analysis frames. Thus, the cross correlation would fail to find any similarity in the signals to compare.

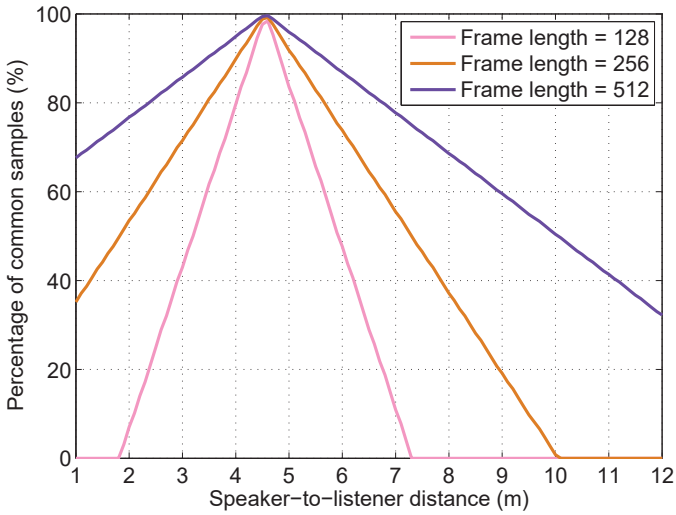


Figure 2.2 – Percentage of common samples between the radio and e.g. the left audio frames, as a function of the speaker-to-listener distance, for a 128-sample (pink), a 256-sample (orange), or a 512-sample (purple) frame size. From [40].

As it is not based on any similarity research in the signals to process, the time delay estimation based on the IACS is suitable in this application. In the LFs, the following relations are assumed to hold for BHAs:

$$\begin{cases} s_L(t) = s_X(t - \Delta_t) \\ s_R(t) = s_X(t - \Delta_t - \delta) \end{cases}, \tag{2.9}$$

where Δ_t denotes the delay between the demodulated audio signal s_X and the signals from the HA microphones s_L and s_R. Computing the cross-spectrum R_L and R_R in both hearing

instruments leads to:

$$\begin{cases} R_L(f) = ||S_X(f)||^2 e^{2\pi f \Delta_t} \\ R_R(f) = ||S_X(f)||^2 e^{2\pi f \Delta_t} e^{2\pi f \delta'} \end{cases} \tag{2.10}$$

where S_X is the Fourier Transform of s_X. The complex values of R_L and R_R at a certain frequency f_0 can be exchanged between the 2 devices, and the ITD can be recovered by deriving the ratio between these 2 values, i.e.:

$$\delta = \frac{\angle\left(\frac{R_R(f_0)}{R_L(f_0)}\right)}{2\pi f_0}. \tag{2.11}$$

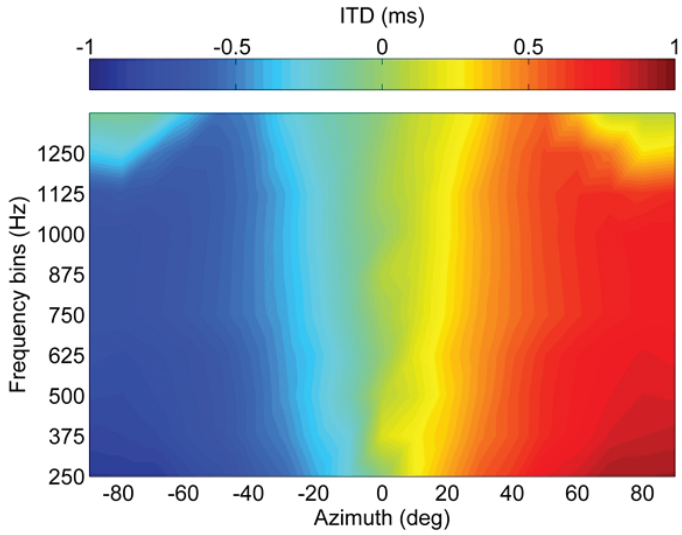


Figure 2.3 – The ITD extracted from the IPD at different frequency bins of a 32-point FFT, as a function of the azimuth. Taken from [40].

Figure 2.3 represents the ITD extracted from the IPD (Equation 2.11) for some of the different frequency bins of a 32-point FFT. The azimuth range is between -90° and 90° (10° steps). The measurements were done in an anechoic environment. The stimulus was a white noise that was captured by 2 microphones behind the ears of a knowles electronic manikin for acoustic research (KEMAR). The 250, 375 and 500 Hz center frequencies are the ones with the greater range of IPD values. Hence they were chosen for the estimation of the IPD. Additionally, these frequencies cover the typical pitch of the voice. Note that the phase ambiguity that is evoked in Appendix A.2.1 appears at the 1250-Hz bin. In practice, the reported algorithm derives a 3-bin FFT of the signals s_X and s_L in the left device, and the FFT of s_X and s_R in the right device. A prior downsampling by a factor of 4 is applied on the analysis frames. This decimation is performed to speed up the computation of the FFT, while preserving the LF resolution.

The phase is then extracted and exchanged between both HAs. Finally, the phase values are subtracted in order to yield 3 observed IPD values:

$$\tilde{\boldsymbol{\varphi}} = \begin{pmatrix} \tilde{\varphi}(f_1) & \tilde{\varphi}(f_2) & \tilde{\varphi}(f_3) \end{pmatrix}^T, \tag{2.12}$$

where ^{T} denotes the transpose operator.

The observation vector $\tilde{\boldsymbol{\varphi}}$ is compared with some theoretical IPD values calculated from the sine law (Equation A.2):

$$\varphi(\theta, f) = \frac{2\pi f a}{c} \sin\theta, \tag{2.13}$$

where θ is the DOA ($\theta \in [-\pi : \pi]$), c is the speed of sound, and a is the distance between the 2 microphones modeling the ear entrances, taken greater than the average head diameter, since the model ignores the curved path around the head. A collection of theoretical IPD values for N azimuths θ are computed at the 3 considered frequencies, and gathered in the matrix $\boldsymbol{\Phi}$:

$$\boldsymbol{\Phi} = \begin{pmatrix} \varphi(\theta_1, f_1) & \varphi(\theta_2, f_1) & \cdots & \varphi(\theta_N, f_1) \\ \varphi(\theta_1, f_2) & \varphi(\theta_2, f_2) & \cdots & \varphi(\theta_N, f_2) \\ \varphi(\theta_1, f_3) & \varphi(\theta_2, f_3) & \cdots & \varphi(\theta_N, f_3) \end{pmatrix}. \tag{2.14}$$

Thus, the observation vector $\tilde{\boldsymbol{\varphi}}$ can be used to estimate the current source location $\tilde{\theta}$, i.e. the azimuth angle such that:

$$\tilde{\theta} = \underset{\theta_{j, j \in \mathbb{N}_N^*}}{\arg\min}\, \epsilon(\theta_j), \tag{2.15}$$

where $\epsilon(\theta_j)$ is the error between the observed and theoretical IPD values for the collection of tested angles θ_j. Since the phase of the cross-spectrum is extracted modulo 2π (the so-called phase wrapping, see e.g. Fumitake *et al.* [75] for more details), it makes no sense to compute an Euclidian distance to the model, which would be biased by those 2π shifts. Instead, a distance criterion based on a sine function is preferred and defined as follows:

$$\begin{aligned} \epsilon(\theta_j) = & \frac{1}{3} \sum_{l=1}^{3} \sin^2(\boldsymbol{\Phi}(l, j) - \tilde{\boldsymbol{\varphi}}(l)), \\ & \text{for } j = 1, 2, ..., N. \end{aligned} \tag{2.16}$$

A way to counteract the detrimental effect of noise and reverberation is to reject all the frames that lead to an IPD pattern significantly different from the model. This might indicate the presence of multiple reflections in the signal, as well as interfering noise that distorts the

pattern. In order to select a certain frame, it is ensured that the smallest error among the collection of angles θ_j is lower than a defined threshold ξ, i.e.:

$$\min_{\theta_{j,j\in\mathbb{N}_N^*}} \epsilon(\theta_j) < \xi. \tag{2.17}$$

A summary of the process for the IPD computation is shown on Figure 2.4.

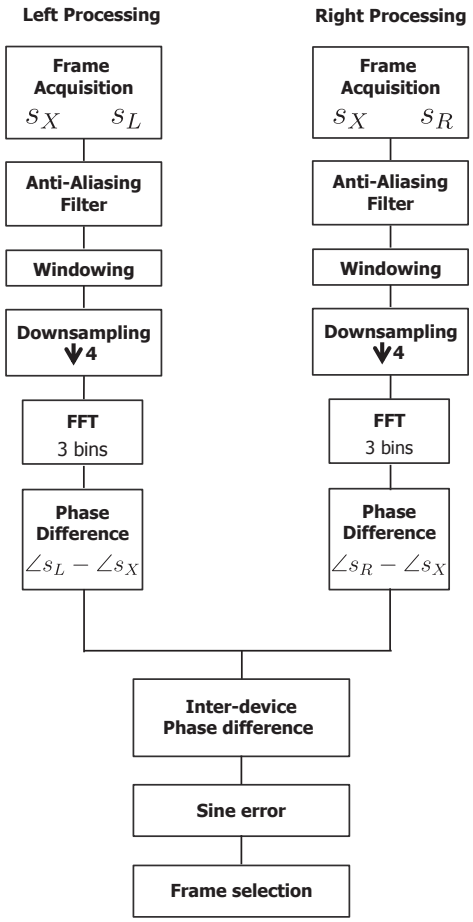


Figure 2.4 – Block diagram of the entire algorithm for the IPD computation. From [40].

2.2.2 Side estimation

While the IPD is used as a fine cue, the ILD and the *received signal strength indication difference* (RSSID) do not manage to provide an accurate localization, for the reasons explained below. Therefore, both cues are exploited so as to bring a rough localization, in the form of a left/center/right output. Their processing is the object of the current part.

Interaural level difference

In the BLAs reported in part 2.1.2, the ILD is rarely used. Indeed, this cue is more sensitive to noise and reverberation than the IPD, and thus less reliable. Lim and Duda [139] estimate the ILD dividing the absolute value of the zero-lag autocorrelation of the left signal by the absolute value of the zero-lag autocorrelation of the right signal. The calculation is performed in sub-bands via a bank of 45 band-pass filters, resulting in 45 ILD estimations. Then, these values are compared to a set of reference values. Unfortunately, no information is given on the frame duration. The reported localization error in the horizontal plane (2°) corresponds to measurements of a single pulse in an anechoic chamber, i.e. far from real-world conditions. Raspaud *et al.* partly resort to the ILD in their BLA, deriving the ratio of amplitude between the left and right signals. They use the following formula to infer the location of the incidence direction:

$$\mathrm{ILD}(f) = b(f)\sin\theta, \tag{2.18}$$

where $b(f)$ is a sequence of scaling coefficients computed from the HRTFs of 45 subjects. Testing their algorithm on sounds of musical instruments in an optimal acoustic environment, they notice that their ILD range remains constant as soon as the source-to-microphone distance is higher than 1.5 m.

Diffuse noises tend to decrease the ILD range, because they provide a constant and similar SPL to both HAs. The lower the SNR, the lower the range. Additionally, local and close noises (e.g. coughs, chats...) can easily yield overestimated ILD values, especially if the ILD is computed on short frames. The same problem occurs if the listener is near a reflective surface (wall, window...). The SPL is artificially amplified at one side and the resulting ILD estimation is skewed. On the contrary, the range of the ILD diminishes with the decrease of the DRR (increase of the speaker-to-listener distance). All these drawbacks prevent the use of the ILD as a precise localization cue.

In the reported BLA, the ILD is estimated in real time in order to state whether the sound source is on the left or on the right side relative to the listener. A positive ILD corresponds to a speaker located on the left, while a negative value is interpreted as a speaker located on the right. When the current estimation is close to 0, the value is discarded as it may correspond either to a sound coming from the left or from the right, considering a certain margin of error. In this case, the side is set to an "unknown" status. Figure 2.5 depicts some measurements of the ILD in an anechoic chamber with a KEMAR wearing 2 microphones behind the ears, for a sound source on the right side (negative ILD). Figure 2.5A shows the amplitude of the ILD as a function of the azimuth and in different frequency bins. As indicated in Appendix A.2.1, the ILD is close to 0 in the LFs, because the head shadow effect does not occur. The 2-4 kHz and 4.5-6 kHz frequency bands provide the highest ILD range. In both cases the ILD appears not to be a one-to-one function of the azimuth (e.g. same ILD values at -60° and -80° in the HFs).

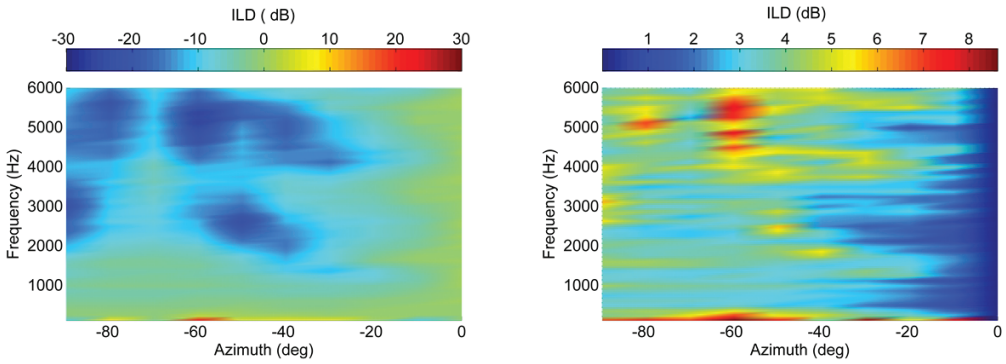


Figure 2.5 – Average ILD measured in an anechoic chamber (A), and the corresponding standard deviation (B), for a sound source on the right. Taken from [40].

That is a rationale to not use the ILD as a way to achieve a precise localization, but only as a side indicator. Figure 2.5B displays the *standard deviation* (SD) of the ILD values obtained for all azimuths and frequencies. The upper bandwidth is affected by a strong SD. Therefore the 2-4 kHz frequency band is preferred. The estimation of the ILD is thus performed after a band-pass filtering operation. The same trend was observed with a speech stimulus, although it appeared to be less precise.

As mentioned, the short-time observations of the ILD exhibit some important variations that prevent from getting a stable output. ILD values computed on several tens of milliseconds are significantly steadier. In order to avoid storing successive audio frames in a buffer (i.e. a waste of physical memory), it is preferred to accumulate the energy of each consecutive frame until the ILD can be derived. The energy E of the frame k is calculated as follows:

$$E_L(k) = E_L(k-1) + \sum_{n=1}^{128} |s_L^k[n]|^2, \tag{2.19}$$

where k is the frame index. The relation also holds on the right side, replacing s_L by s_R and E_L by E_R. After a fixed duration, the ILD is estimated computing the ratio between the left and right energies, which requires the interaural exchange of a single value.

Received signal strength indication difference

WMS do not only give access to a clean speech signal from the remote microphone. They also provide the opportunity to analyze the RF transmission from the emitter to the RF receivers in the HAs. The RSSI is an additional spatial cue that is exploited in the localization algorithm. The strength of the RF signal that reaches the left and right RF receivers depends on the location of the speaker, due to the head shadow effect caused by the head on the electromagnetic

waves. Therefore, the RSSID can be viewed as an ILD occurring in the RF domain. Since the RF information is transmitted at about the speed of light, the difference of time of flight between both HAs is far too small to be compared. Fischer [73] has conducted some measurements with a phantom head that reproduces the RF properties of the human head. He reports time differences between 0 and 700 ps for azimuths going from 0 to 90°. Such orders of magnitude require a specific hardware to be detected, which is unrealistic in the context of HAs (too much power consumption and computational cost).

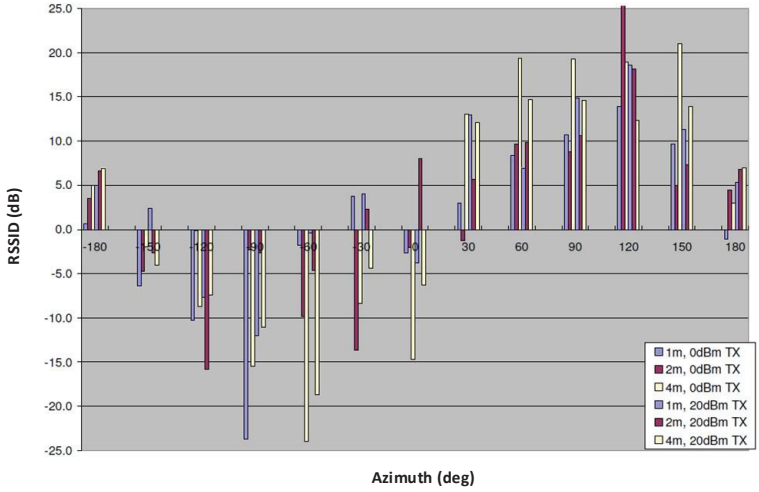
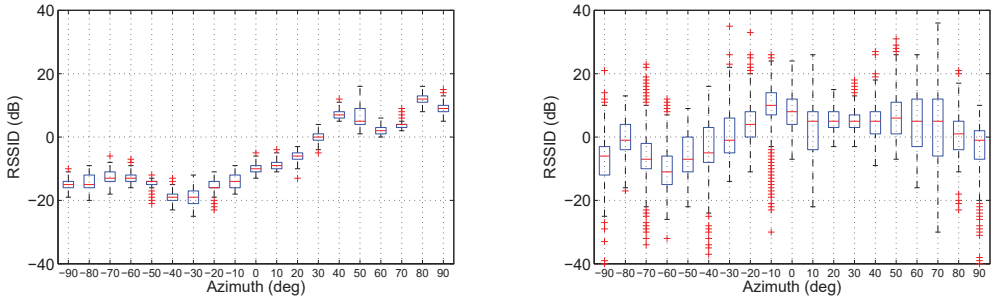


Figure 2.6 – RSSI measurements performed on a head phantom in a dedicated room at different distances and radiation levels. Taken from [206].

The resort to RSSID for localization and distance estimation is already reported in the literature, see e.g. [9, 12, 36, 144, 219]. However, the idea to combine audio-based and RSSID-based localization has not been suggested so far, to the knowledge of the author. This is why Phonak patented the concept in the field of WMS in 2011 [183]. At the early stages of the research, the reliability of a concrete RSSID use has been evaluated. Figure 2.6 shows the results from a campaign of RSSID measurements done with a head phantom at 2.45 GHz in a dedicated room (i.e. far away from reflective surfaces and RF interferences). The measurements were performed at 2 radiation levels (0 and 20 dBm) and for 3 increasing distances between the emitter and RF receivers (1, 2 and 4 m). Each bar corresponds to the mean of 3 RSSID values of 100 ms each. At first glance, the RSSID follows a nice trend, being negative for azimuths on the right and positive for azimuths on the left. No significant difference can be noticed between the 2 radiation powers. The values seem to depend on the distance, but in a random-like way. Some wrong RSSID values are located at azimuths between ± 30°. From these outcomes, one can conclude that the RSSID looks interesting to estimate the side of the speaker but cannot be exploited to determine its precise position . As for the ILD, it plays the role of a side indicator. A decision threshold for the “unknown” status has been fixed as well, i.e. no information about

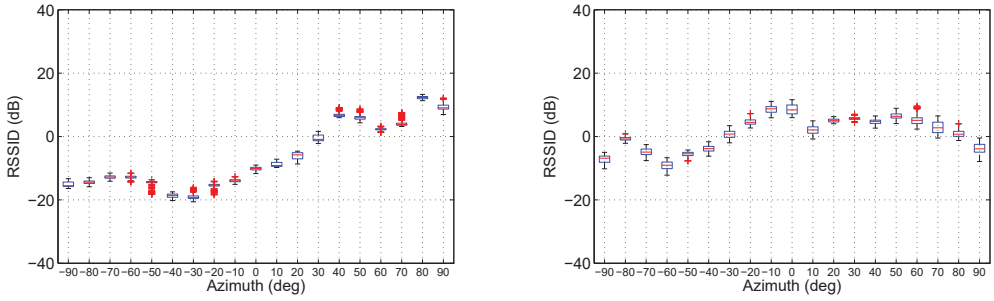
the side is given in the corresponding range of values.

In the targeted hardware, a new packet of RF-transmitted information reaches the HAs each 4 ms. This means that 250 RSSID values can be computed every second. Figure 2.7A depicts the distribution of several hundreds of RSSID values computed for various azimuths in the FHP in a RF-anechoic environment. The RF receivers are plugged on a pair of HAs worn by a KEMAR. The head of the latter has been filled with electromagnetic (activated carbon) absorbers, so as to approximate the absorption property of the human head. An offset of 10 dB can be noticed at 0°, which might be due to an asymmetry of the setup and/or a different sensitivity of the antennas. Anyway, one can see that the distribution is well-shaped with only little outsiders. On the contrary, the outcomes from the same setup in a typical classroom are non-exploitable (Figure 2.7B), due to the presence of too many outsiders. Nevertheless, the global shape of the distribution is not totally odd.



(a) (b)

Figure 2.7 – RSSID distributions as a function of the azimuth in a RF-anechoic chamber (A) and in a typical classroom (B).



(a) (b)

Figure 2.8 – RSSID distributions from Figure 2.7 smoothed with a leaky integrator ($\lambda = 0.95$).

It is obvious from the previously reported measurements, that the integration of the RSSID in the localization process requires a smoothing of the values. This can be achieved by the use of a moving average procedure, which outputs at each frame the mean over the last M

values. However this technique demands the storage of M RSSID values. That is why the *leaky integrator* method has been preferred [194]. In this procedure, the current smoothed RSSID output is derived as follows:

$$\overline{\text{RSSID}}[k] = \lambda\overline{\text{RSSID}}[k-1] + (1-\lambda)\text{RSSID}[k], \tag{2.20}$$

where k is the frame index, $\overline{\text{RSSID}}[k-1]$ is the smoothed RSSID computed at the previous frame (i.e. the only value to be stored), RSSID$[k]$ is the current "raw" RSSID value, and λ is the leaky integrator coefficient. This technique has been applied on the previous measurements and the results are displayed on Figure 2.8. Both distributions show to be quite less spread, but this is at the cost of a lower range of values. Still, it remains reversed RSSID outputs at 10 and 20° .

One of the 3 major advantages of the RSSID cue is that it is available even when the speaker does not talk, because the emitter keeps on transmitting data corresponding to silent frames. This allows to update the localization during non-speech periods. Another advantage comes from the fact that materials do not have the same behavior with reflected electromagnetic and acoustic waves. Indeed, if the ILD is degraded in a certain configuration, there is a possibility that the RSSID is not, and conversely. Thus, it is possible to correct the possible misleading information coming from one cue thanks to the other one. The way this is performed is described in the next part. The combined use of the ILD and RSSID is another technique to reduce the influence of noise and reverberation. Finally, this information is free and inexpensive, as recalled by Cheng *et al.* [36].

2.3 Localization & tracking

Here is discussed how the BLA exploits the information from the various spatial cues to end up with an estimated location of the speaker. The principle of the tracking procedure is then detailed.

2.3.1 Localization

Once all spatial cues are available, the algorithm has to merge the different information to infer the location of the speaker. The spatial resolution of the reported BLA has been set to 5 spatial sectors covering the FHP, as shown on Figure 2.9. Their angular span is around 35°, although the sectors do not strictly have the same angular aperture. This resolution has been chosen according to 3 main reasons. First, the algorithm is designed for applications where the visual cue of listeners is available and predominant. The plasticity of the brain, evoked in Appendix B.2.1, is expected to enable the matching of a rough acoustic resolution with a precise visual stimulation. Second, the end-users are HI subjects, whose localization performance has been reported as partially degraded. A fine accuracy may not be required nor

useful for such listeners. Finally, a coarse resolution helps increase stability and diminishes the risk of strong localization errors.

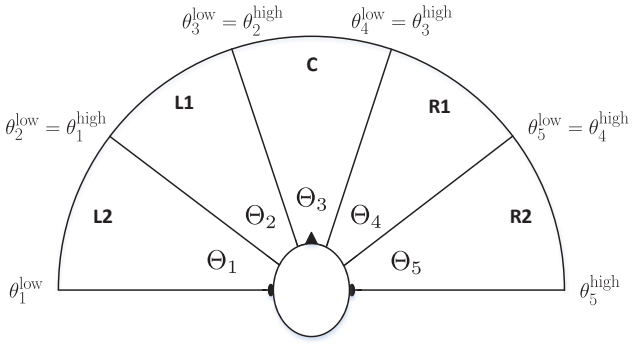


Figure 2.9 – Spatial resolution of the reported BLA with 5 sectors in the FHP. Taken from [52].

As soon as the angular errors between the observed and theoretical IPD values (Equation 2.16) pass the selection step (Equation 2.17), they are gathered into sectors as follows:

$$\sigma(\Theta_i) = \frac{1}{M_i} \sum_{\theta_j=\theta_i^{low}}^{\theta_i^{high}} \epsilon(\theta_j) \qquad (2.21)$$
$$\text{for } i = 1,2,...,5, \text{and } j = 1,2,...,M, M \geq 5,$$

where $\sigma(\Theta_i)$ denotes the accumulated error of the sector Θ_i, θ_i^{low} and θ_i^{high} are the low and high azimuth boundaries of the corresponding sector (as shown on Figure 2.9), and M_i denotes the number of discretized angle that compose the sector Θ_i. The 5 resulting values can be converted into probabilities for the source to be in the 5 different sectors. Equation 2.16 has been defined such that $\sigma(\Theta_i)$ is smaller than 1, preventing the IPD-based probability $p(\Theta_i)$ to be in the sector Θ_i from being less than 1. $p(\Theta_i)$ is defined as follows:

$$p(\Theta_i) = \frac{1-\sigma(\Theta_i)}{\sum_{j=1}^{5} 1-\sigma(\Theta_j)} \qquad (2.22)$$
$$\text{for } i = 1,2,...,5,$$

Note that the notion of "probability" used here does not correspond to the mathematical definition of the probability *stricto sensu*. It must be rather interpreted as an indicator of the likelihood of the source to be in a given sector [52]. Equation 2.22 is such that a large difference between the observation and model for a given sector leads to a low probability of being in the corresponding sector, and conversely. The probabilities from the IPD block are averaged over a certain number of frames before being mixed with the ILD and RSSID contributions.

The side information coming from the ILD and RSSID is taken into account by applying some weightings on the computed probabilities. The probabilities of being in the 2 sectors of the current side are emphasized, while the probabilities of being in the 2 sectors of the opposite side are lessened. For instance, if the ILD indicates that the source is located on the left, the probabilities of being in the sectors L1 and L2 (see Figure 2.9) are increased, while the probabilities to be in the sectors R1 and R2 are decreased. The probability of the central sector C is never affected by the weighting operation. Mathematically, this can be formulated as follows:

$$\begin{aligned} S_W(\Theta_i) &= W_{\text{ILD}}(\Theta_i) W_{\text{RSSID}}(\Theta_i) p(\Theta_i) \\ &\text{for } i = 1,2,...,5, \end{aligned} \tag{2.23}$$

where $S_W(\Theta_i)$ denotes the weighted score associated with the sector Θ_i, and W_{ILD} and W_{RSSID} are the weightings applied on the IPD-based probabilities $p(\Theta_i)$. For example, if the ILD and RSSID better match a speaker located on the left, the weights are:

$$W_{\text{ILD}}(\Theta_i) = W_{\text{RSSID}}(\Theta_i) = \begin{cases} \nu \text{ for i} = 1,2 \text{ (left sectors)} \\ 1 \text{ for i} = 3 \text{ (central sector)} \\ \dfrac{1}{\nu} \text{ for i} = 4,5 \text{ (right sectors)} \end{cases}, \tag{2.24}$$

with $\nu > 1$. The weightings of the ILD and RSSID are defined in a similar manner, so that the contributions of the 2 cancel each other in case of a contradictory side indication. This helps reject erroneous cue estimations, as no cue appeared to be better than the other. If the output is set to “unknown”, all weights are equal to 1.

2.3.2 Tracking

The last step of the localization processing consists in a tracking procedure, which has been implemented in order to enhance the system stability. The general idea of tracking is to take into consideration the previous estimated locations when determining the current one. The tracking model developed in the presented BLA is a probabilistic network where the 5 spatial sectors are represented by 5 nodes connected with arrows, as depicted on Figure 2.10. Each arrow corresponds to a probability to go from a sector to another. Every node is also connected to itself by an arrow representing the probability to stay in the current sector. Given that the source is currently located in the sector Θ_j, the transition probabilities are such that:

$$\begin{aligned} &\sum_{i=1}^{5} p(\Theta_i|\Theta_j) = 1 \\ &\text{for } j = 1,2,...,5, \end{aligned} \tag{2.25}$$

where $p(\Theta_i|\Theta_j)$ denotes the probability to move to the sector Θ_i knowing that the sound source was previously located in the sector Θ_j. These transition probabilities act as weightings that emphasize or lessen the current scores associated with each sector. Their values have been set empirically. Since it is unlikely that the speaker suddenly moves e.g. from the sector L2 to R2, the transition probability modeling this path is set to a small value. If the current observation predicts so, either it is an error and it is discarded by the tracking procedure, or it is true and several iterations are required before detecting the sector change. The integration of the tracking procedure thus gives a certain inertia to the BLA. The research of an adequate tradeoff between accuracy and speed is the topic of Chapter 3.

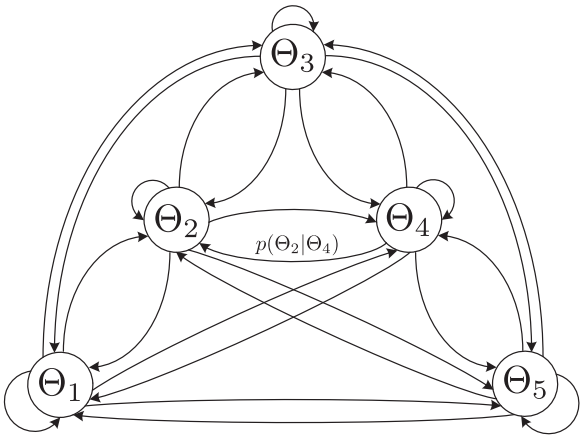


Figure 2.10 – Probabilistic network governing the tracking procedure of the BLA. Taken from [52].

Let j the index of the estimated sector for the previous frame, $j \in \mathbb{N}_5^*$. The transition probabilities are applied as follows:

$$S_W(\Theta_i|\Theta_j) = p(\Theta_i|\Theta_j) S_W(\Theta_i) \quad \text{for } i = 1,2,...,5, \tag{2.26}$$

where $S_W(\Theta_i)$ is computed from Equation 2.23, and $S(\Theta_i|\Theta_j)$ denotes the weighted score of moving to the sector Θ_i knowing that the source was previously located in the sector Θ_j. The location estimation of the frame k, Θ_k (spatial sector), for the current observations of IPD, ILD and RSSID, considering the previous location of the source Θ_j, is:

$$\Theta_k = \underset{\Theta_{i,i\in\mathbb{N}_5^*}}{\arg\max}\, S_W(\Theta_i|\Theta_j). \tag{2.27}$$

2.4 Additional features

In order to improve the performance of the BLA in terms of accuracy, 2 other signal processing features have been introduced in the process. This part presents both of them and highlights their respective contributions to the algorithm.

2.4.1 Voice activity detection

A *voice activity detector* (VAD) is an algorithm that detects the presence of speech in an input signal. VADs are present in numerous fields of speech processing, e.g. speech coding, speech recognition or speech enhancement [198, Chap. 1]. There are several rationales for the use of VADs: reduce the bitrate of a transmission channel, perform noise reduction (as explained in Chapter 1.3), improve the efficiency of speech coding... Many approaches exist to achieve voice activity detection. Bio-inspired methods are based on periodicity cues (pitch detection, modulation depth, spectrum analysis...) as done in the AS, but other techniques have been developed as well, resting upon e.g. the zero-crossing rate, the entropy content, or the cepstral information [198, 210, Chap. 1]. The typical output of a VAD is shown on Figure 2.11A. The boolean output is set to 1 as soon as the speaker talks. During silences or breathings the boolean equals 0.

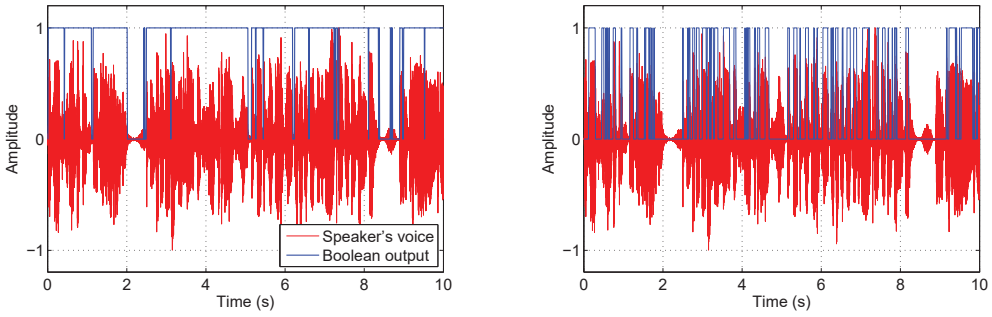


Figure 2.11 – Input speech signal (red) and the boolean output (either 0 or 1) of a VAD (blue), from a common VAD (A) and from the VAD implemented in the BLA (B). Taken from [40].

In the presented BLA, the use of a VAD is recommended to decrease the processing power and enhance the accuracy. The clean speech signal from the radio signal is a good candidate to be processed by a VAD. Performing the localization procedure when the speaker does not talk would lead to false determined locations, since only the background noise would feed the algorithm. Moreover, as the BLA is primarily based on phase difference estimations in the LFs, the speech components that do not have enough energy in the LFs (e.g. fricatives) must not be included either. Therefore, the VAD implemented in the BLA rather acts as a voiced-phoneme activity detector, as shown on Figure 2.11B, where the VAD is true for some speech portions only. It rests upon a combinatory logic-based procedure between some conditions on the

energy computed in 2 sub-bands [49]. The VAD is the first stage of the BLA, and transmits the information of all the frames that must stay unprocessed to the further blocks. In this way, the computationally expensive instructions of the algorithm are not run at each new frame.

2.4.2 Estimation of the environment quality

Although the BLA incorporates some strategies to counteract the detrimental effects of noise and reverberation, the performance shows to worsen in nasty conditions, especially when the speaker-to-listener distance exceeds 4 or 5 m. The acoustic signals captured by the HA microphones are more and more degraded and it becomes difficult to extract information about the targeted speech.

With the knowledge of the fixed processing delays in the emitter and in the HAs, it would be possible to estimate the distance between the speaker and listener. This was one of the optional objectives stated in Chapter 1.4.3. However, this would require the computation of a cross-correlation, which is inconceivable for the reasons indicated in part 2.2.1. Moreover, the effect of the distance on the algorithm performance highly depends on the environment, i.e. the BLA can perform well in favourable conditions up to e.g. 6 m, while a distance of 3 m would be sufficient to deteriorate the outcomes in a reverberant room. That is why the notion of *intermodal coherence* was preferred.

In Chapter 1.3, it has been reported that the derivation of the IC is as new tool to design noise reduction algorithms in BHAs. Indeed, this cue can be viewed as a sound quality estimator. In the reported BLA, an analysis of the resemblance between the body-worn and HA microphone signals can provide an interesting indicator about the conditions the algorithm currently copes with. It has been referred to the *coherence estimation* (CE) block in the following. In the left HA, it is computed as follows [69, Chap. 5]:

$$\mathrm{IC}[K] = \max_{m} \frac{\overline{s}_L^{k\rightarrow k+4}[n]\overline{s}_X^{k\rightarrow k+4}[n+m]}{\sqrt{\overline{s}_L^{k\rightarrow k+4^2}[n]\overline{s}_X^{k\rightarrow k+4^2}[n+m]}}, \quad \text{for } K = 1,6,11..., \tag{2.28}$$

where $\overline{s}_L$ and $\overline{s}_X$ denote the means of s_L and s_X, k is the frame index, K is the index corresponding to the output rate of the CE block, n is the sample index and m is the varying sample shift index. The same relation holds in the right device, replacing s_L by s_R. The CE block faces the same issues as reported in part 2.2.1. It requires to buffer 5 consecutive frames, as described by the notation $k \rightarrow k+4$ in Equation 2.28. Because of that, it has been decided to decimate all the signals by a factor of 7 in the buffers, so as to reduce the buffer size down to 92 samples, and considerably accelerate the calculation of the intermodal coherence. In order to save time, no anti-aliasing low-pass filter is included before the downsampling. Interestingly, the outcomes from the resulted intermodal coherence are only dimly affected by this operation, as shown on Figure 2.12. The main rationale for this is that both s_L (or s_R) and s_X are subsampled in the same way. Figure 2.12A displays the results from the CE block obtained with a speaker

located successively at 3 and 6 m from the listener in an auditorium. The blue solid line corresponds to the consecutive "raw" values of intermodal coherence, while the red solid line is the smoothed output. As expected, the coherence decreases (from around 0.5 down to 0.35) when the speaker-to-listener distance changes from 3 to 6 m, and the speaker motion is clearly visible (green boxes). Figure 2.12B depicts the output of the CE block after the decimation. The outcomes are quite similar to Figure 2.12A. The general trend is the same, only the range decreases from 0.24 down to 0.2. This is due to 2 main factors: the artifacts yielded by the absence of anti-aliasing filter, and the fact that the reduction of the bandwidth is equivalent to a coherence computation in the LFs, where the IC is known to present a smaller range. Anyway, the results support the use of decimation in the intermodal coherence estimation procedure.

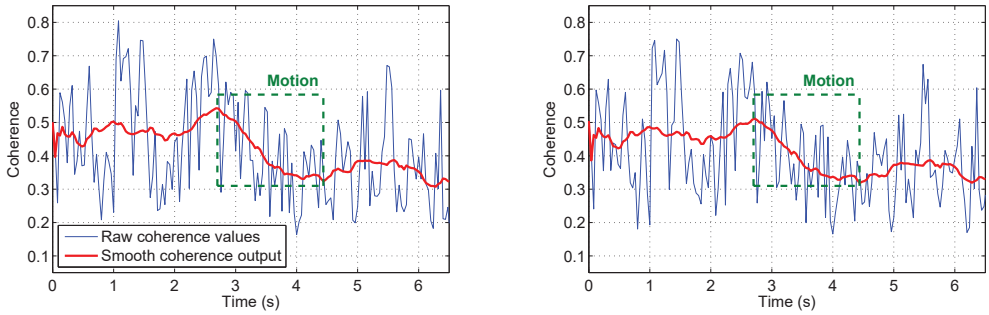


Figure 2.12 – Intermodal coherence corresponding to a speaker located successively at 3 and 6 m from the listener in an auditorium, computed with the original signals (A) or the downsampled signals by a factor of 7 (B). In blue, the successive "raw values" and in red the smoothed ones. The time segments corresponding to the motions are indicated by the green boxes.

Practically, the CE block computes the intermodal coherence over 5 downsampled frames, if all corresponding VAD outputs are true. It alternates a comparison between the signals s_X and s_L on the 5 first frames, then s_X and s_R on the next 5 frames, and so on. This allows to take into consideration both HA audio signals, without deriving the coherence in the 2 devices simultaneously. A moving average over a certain number of last values is finally performed to ensure a stable output. This means that only 1 value needs to be exchanged between both apparatus every 10 frames. The total storage on the device that computes the smoothed coherence is composed of 213 values (i.e. $2 \times 92 + 29$), instead of 1280 values if no decimation was performed. That is, 84% of memory is saved thanks to this procedure.

A threshold of minimum acceptable coherence has been fixed. Below this value, the BLA stops working and no spatialization is performed either. This is to avoid an erroneous estimated location of the speaker. With small values of IC, the algorithm has not shown to end up with an adequate localization in 5 spatial sectors (unstable and possibly wrong output). Therefore, the resolution is set to 3 sectors (left/center/right) in this case. Then, the 5-sector resolution is

available. A typical output from the CE block is depicted on Figure 2.13. It corresponds to a situation where the speaker moves from 6 to 3 m relative to the listener, in a listening room (treated with carpet and absorbers on 3 walls, while large windows cover the last one) and in an auditorium. The speaker was a real person speaking in the environment and wearing the emitter device (Phonak Roger inspiro) with its microphone. The sound signal was captured with 2 HAs (Phonak Naida IX SP) worn by a KEMAR and collected/processed by a Simulink model of the algorithm. Like on Figure 2.12, the coherence augments with the diminishing distance. Also, it depends on the quality of the environment, going from 0.38 to 0.64 in the favourable listening room, while it goes from 0.32 to 0.48 in the auditorium. The speaker moved at a constant speed of 0.4 km/h. The resolution toggles from 3 to 5 sectors at a certain time, which will be shorter in the case of the listening room than in the auditorium. This is exactly what is required from the CE block.

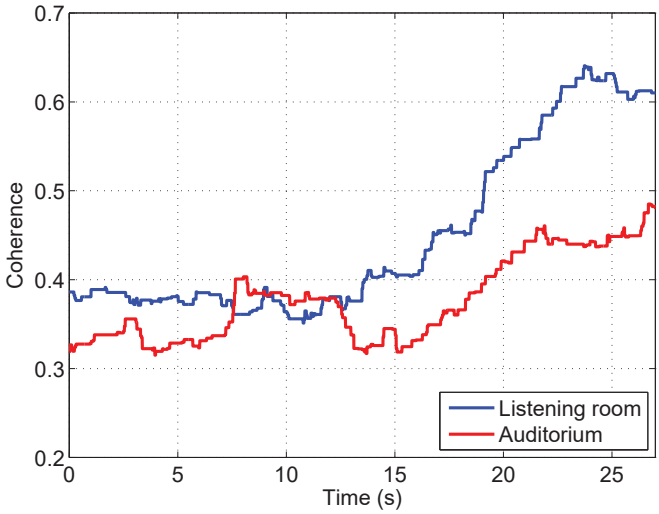


Figure 2.13 – Intermodal coherence (smoothed over 30 frames) resulted from the speaker motion from 6 to 3 m relative to the listener at a constant speed of 0.4 km/h. Measurements done in a listening room (blue solid line) and in an auditorium (red solid line).

2.5 Conclusions

This chapter has introduced the 2 main contributions related to the localization algorithm:

1. The integration and combination of 2 types of localization cues, namely the IPD and ILD from the acoustic propagation, and the RSSID from the electromagnetic propagation,
2. The development of a BLA working in real time, while respecting the technical constraints demanded by WMS.

Figure 2.14 present the flow chart of the whole BLA, including all the reported processes:

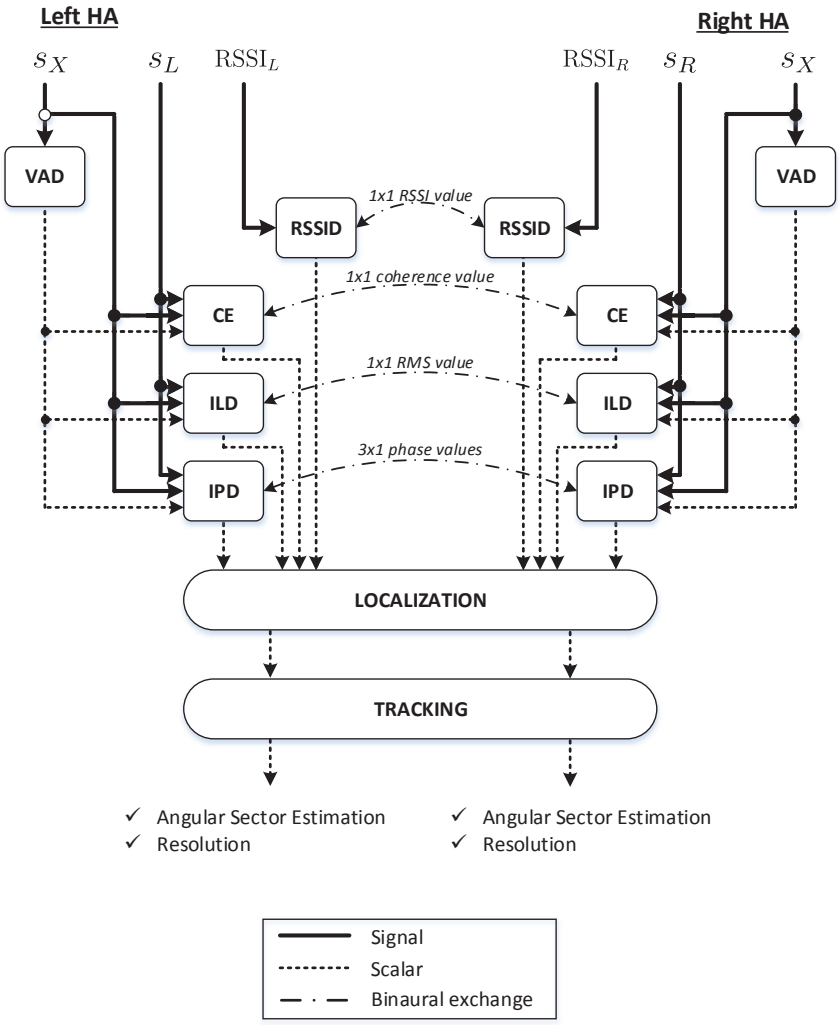


Figure 2.14 – Block diagram of the entire BLA.

- The VAD block, which determines which input frames of the audio signals must be processed, and transmits this information in the next 3 blocks,
- The CE block, which monitors the spatial resolution of the algorithm by determining the quality of the acoustic environment,
- The ILD block, which is in charge of delivering information about the speaker's side location from the difference of SPLs between both HAs,
- The IPD block, which is the core of the localization procedure. It estimates the phase difference between both devices and compares them to some reference values, so as to output an error value in each azimuth between -90° to 90°,

- The RSSID block, which also plays the role of a side indicator by analysing the strength of the electromagnetic signals reaching the 2 RF receivers,
- The Localization block, which is fed by the information coming from the IPD, ILD, RSSID and CE blocks. Its output is a set of 3 or 5 probabilities for the speaker to be in one of each spatial sector,
- The Tracking block, which takes into account the previous estimated location of the speaker to avoid abrupt changes of position.

The kinds of communications that link all these blocks are displayed as well. They can be signals, data, or wireless contents. The final output of the BLA gives the information about the speaker's location in one of the 3 or 5 spatial sectors, and the current resolution of the algorithm. The way how it has been optimized is the object of Chapter 3.

The respect of the technical constraints has been possible thanks to a methodic approach, yielding the following main strategies:

- The limitation of the binaural exchanges. In particular, the absence of transmission of audio frames between both HAs. Instead, only some single data go from one device to the other in the IPD, ILD, RSSID and CE blocks,
- The limitation of the processing power. Some selection procedures are used to avoid the processing of the most costly calculations when they are useless. Another example is the direct derivation of the 3 bins of interest in the IPD block, instead of computing a full 32-point FFT,
- The limitation of the memory usage. Low-order filters have been privileged (e.g. 5th order filters for the VAD, 6th order filter in the IPD block...). The resort to the leaky integrator method in the RSSID block, and the accumulation of the energy in the ILD block are other examples. Also, the BLA uses a mathematical model of the head that can be calculated each time it is required, instead of being stored. Finally, a last strategy is the decimation of the signals. It is applied as often as possible, to limit the buffer size.

The respect of those constraints have not prevented from elaborating some efficient knacks to ensure some good performance of the BLA in adverse environments. Here is a list of the major strategies developed to cope with noise and reverberation:

- The introduction of a VAD to reject all potential undesired frames (the silent sequences and some consonants in the speech signal),
- The frame selection procedure in the IPD block,
- The combined use of the ILD and RSSID, which is a powerful technique, since situations where both cues are disrupted are less frequent,

- The resort to intermodal coherence estimation to detect nasty acoustic conditions,
- The spatial resolution limited to 3 or 5 sectors,
- The introduction of a tracking procedure, to avoid instabilities in the localization.

The access to a clean signal coming from the remote microphone is a strong advantage provided by WMS, as well as the availability of a non-acoustic spatial cue. The RSSID was hard to integrate in the algorithm, because of a certain limited experience with electromagnetic propagation and antenna theory, and the requirements of a specific hardware of the intended prototype to get it. However, it brings a strong advantage to perform localization in adverse acoustic environments and keep an update of the speaker's location during silences. Nevertheless, all the reported strategies aiming to preserve an accurate localization under complex conditions give a significant inertia to the BLA, because almost all blocks are based on some average information over a certain time. Also, lots of parameters are available in the different blocks to adjust the tradeoff between the accuracy and speed of the algorithm. This is the object of Chapter 3, that reports the performance of the final version of the BLA.

3 Optimization and evaluation of the localization algorithm

This chapter first investigates the sensitivity of the localization algorithm to its numerous parameters. The objective of part 3.1 is to improve and optimize the performance of the BLA. One has to define some scores on which the optimization needs to be performed. In the context of localization algorithms, it is current to resort to 2 antagonist quantities that are the accuracy (to be maximized) and the reaction time (to be minimized). Then, the optimization requires the acquisition of an extensive amount of "real-world" data. This allows to compute the initial performance, select the most sensitive parameters, and perform their fine tuning. To do so, some specific optimization methods are applied. Part 3.2 details the results that are obtained in the different parts of the algorithm, including the IPD, the side estimation (ILD/RSSID) and the VAD blocks. Finally, the overall pre- and post-optimization performance are presented. The conclusions of the chapter are drawn in part 3.3.

3.1 Optimization

This first part is dedicated to the optimization procedure of the BLA. It introduces the scores to optimize, the acquisition of a real-world database to perform the computations, and the methods implemented to end up with an optimized version of the algorithm.

3.1.1 Score definition

The optimization of the BLA is based on 2 different, and somehow opposite, scores. The first one is related to the accuracy of the algorithm (i.e. how often does the algorithm locate the speaker in the adequate sector?). The second represents the reactivity (reaction time) of the BLA (i.e. what is the time required by the algorithm to reach the right sector?). Both scores are effectively contradictory because a good accuracy usually requires a certain amount of integration time, which reduces the speed of the algorithm. Hence, the objective is to find a tradeoff that yields a satisfying accuracy, while preserving a reasonable reactivity.

Figure 3.1A describes the way the accuracy score is computed. The accuracy A describes

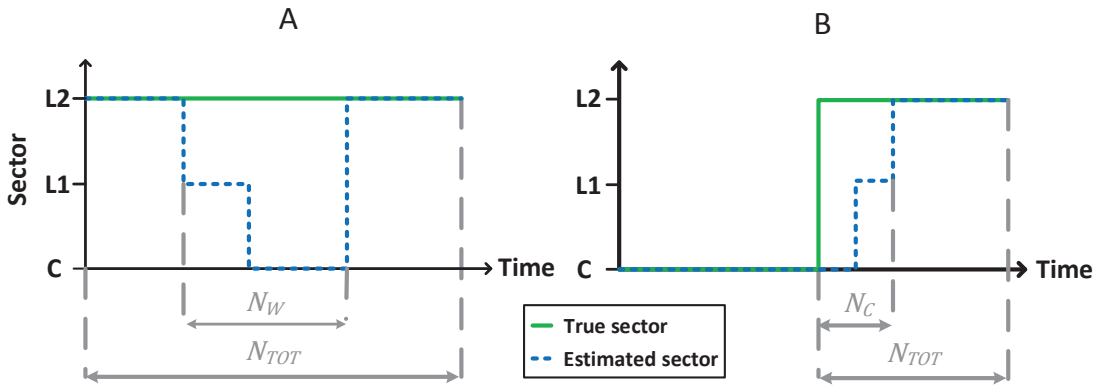


Figure 3.1 – Derivation of the accuracy (A) and the reactivity (B) scores.

how well the estimated spatial sector matches the actual DOA. It is obtained by counting the number of frames leading to a wrong localization during a certain time, and it is expressed as a percentage. A is computed as follows:

$$A(\%) = 100 \times \frac{N_{TOT} - N_W}{N_{TOT}}, \tag{3.1}$$

where N_{TOT} is the total number of analysis frames, and N_W is the number of frames resulting in a erroneous localization (see Figure 3.1A). The accuracy is derived in several azimuths, then averaged into sectors, according to Table 3.1.

Sector	L2	L1	C	R1	R2
Average azimuths	70°, 50°	30°	10°, 0°, -10°	-30°	-50°, -70°

Table 3.1 – Tested and average azimuths for the computation of the accuracy score in the different spatial sectors.

The reactivity R describes how fast the algorithm is able to converge to the correct sector, by counting the number of frames required by the BLA to reach the targeted sector. Figure 3.1B illustrates the principle of the derivation of the reactivity. The postulate is to give priority to the accuracy rather than the reactivity, considering that it would be more unpleasant to listen to a voice spatialized in a wrong location than to have to wait a certain time to get the spatialization updated. Therefore, the determination of the right sector is taken as the target for the calculation of the reactivity, so that the notion of reactivity intrinsically takes into

account the accuracy. It is expressed as a percentage and is calculated as follows:

$$R(\%) = 100 \times \frac{N_{TOT} - N_C}{N_{TOT}}, \tag{3.2}$$

where N_C is the number of analysis frames before the correct sector is output (see Figure 3.1B). The reactivity is evaluated by concatenating 2 stimuli from 2 different spatial sectors, then by counting the convergence time starting from the beginning of the second stimulus. 4 different sector step sizes have been considered, and different transitions are tested and averaged, as summarized in Table 3.2. Note that the term N_{TOT} in Equation 3.2 stands for an arbitrary time reference and has to be taken similar for all configurations. The reactivity is defined in such a way that a high value corresponds to a short reaction time, and conversely.

Step (in number of sectors)	1	2	3	4
Average transitions	-60° → -30° -30° → 0° 0° → 30° 30° → 60°	-60° → 0° -30° → 30° 0° → 60°	-60° → 30° -30° → 60°	-60° → 60°

Table 3.2 – Tested and average transitions for the computation of the reactivity score for the 4 different sector steps considered.

The reactivity is expressed as a percentage, in order to simplify the comparison between the accuracy and reaction time in the following. Indeed, with such a definition, both the accuracy and reactivity must be maximized. However, it is more common to express the reaction time as a time delay that one wants to minimize. This delay can be recovered with the following formula:

$$R_s = N_C \times d_F, \tag{3.3}$$

where R_s is the reaction time (in second), and d_F is the duration of an analysis frame (d_F = 8 ms).

The accuracy and reactivity scores obtained for a certain set of parameters can be viewed as 2 coordinates in a plane (Accuracy × Reactivity), as shown on Figure 3.2. In this plane, the ideal point corresponds to the set of parameters yielding accuracy and reactivity scores of 100%, i.e. the point of which the coordinates are (100;100), depicted in green. The optimization consists in finding the values of the parameters so that the corresponding point is the closest to the ideal one. The Euclidian distance D between the observed point and optimal point permits to

evaluate this distance. It is derived as follows:

$$D = \sqrt{(100 - A)^2 + (100 - R)^2}. \tag{3.4}$$

This distance is depicted on Figure 3.2. The aim of the optimization procedure is to reduce D, considering various acoustic conditions and different types of speakers. Thus, it requires the acquisition of a database, with which the score and distance derivation can be done.

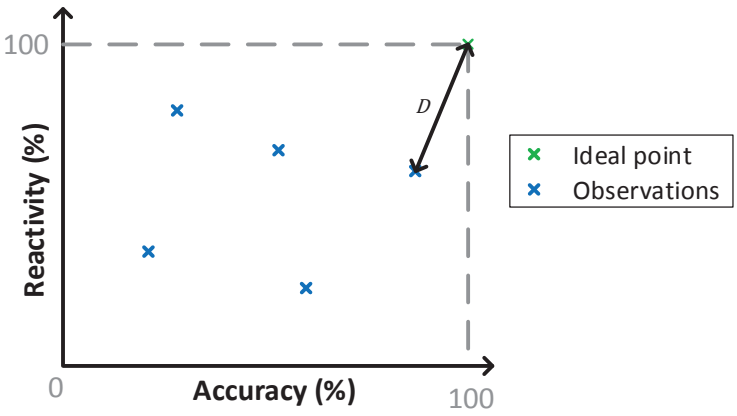


Figure 3.2 – Representation of the scores in the (Accuracy × Reactivity) plane. Some examples of observations are in blue, while the ideal point is in green. The distance D between the observed points and optimal point is represented by the dark double-side arrow. Taken from [43].

3.1.2 Acquisition of acoustic and electromagnetic data

Setup

For the sake of simplicity, the acquisition of the database took place in only 2 rooms: a listening room (i.e. an "optimal" environment) and a typical classroom (i.e. an "adverse" environment). The pictures of Figure 3.3 displays the setup in the listening room (Figure 3.3A) and in the classroom (Figure 3.3B). The global setup is shown on Figure 3.4. 2 manikins were used as the speaker and listener. The speaker was a HATS B&K type 4128. The sound card (Edirol UA101) provided the stimuli to an amplifier (Quad 50E) which was plugged into the artificial mouth of the manikin. A body-worn microphone connected to the RF emitter device (Phonak Roger inspiro) was hitched to its torso. The listener was a KEMAR that was mounted on a turntable. 2 HAs (Phonak Naida IX SP) were worn by the KEMAR, with RF receivers plugged on their DAI. These RF receivers contained an omnidirectional microphone Knowles EK-27263-000 and the RF receiver antenna. Both hearing devices were plugged into a central unit that pre-amplified

the audio signal from the microphones, demodulated the radio signal from the emitter, and extracted the RSSI on both sides. The 3 analog audio signals (s_X, s_L and s_R) were sent to the sound card. The RSSI information was transmitted to the serial port of a PC. The head of the KEMAR had been previously filled with electromagnetic (activated carbon) absorbing material in order to mimic the absorbing property of the human head. The distance between the 2 manikins was constant and set to 4 m for all recordings. Acoustic and electromagnetic absorbers covered the reflective surfaces of the acquisition hardware, so as to limit the possible reflections of the setup that would bias the outcomes.

(a) (b)

Figure 3.3 – Pictures of the measurement setup mounted in the listening room (A) and in the classroom (B).

Stimuli

2 stimuli were used in this experiment. They consisted in a speech signal spoken by a male or a female, in order to study the influence of the speaker gender on the BLA. The male voice was a 18-second sample of the English-spoken sentence available on the EBU SQAM CD [242]. The ISTS (International Speech Test Signal) V1.0 [227] consists of a mixture of 21 female speakers in 6 different languages (American English, Arabic, Chinese, French, German, Spanish). A 18-second excerpt was chosen as the female voice stimulus. The sampling rate was 48 kHz, and reduced to 16 kHz in post-processing. The SPL measured at the position of the center of the KEMAR head was such that the SNR in both environments was equal to 10 dB.

Procedure

In both environments, the KEMAR, mounted on a turntable, was rotated from -90° to 90° by steps of 10°. For each position, the 2 stimuli were successively emitted by the HATS and recorded. 3 consecutive sets of data were acquired in the classroom, and 2 sets were acquired

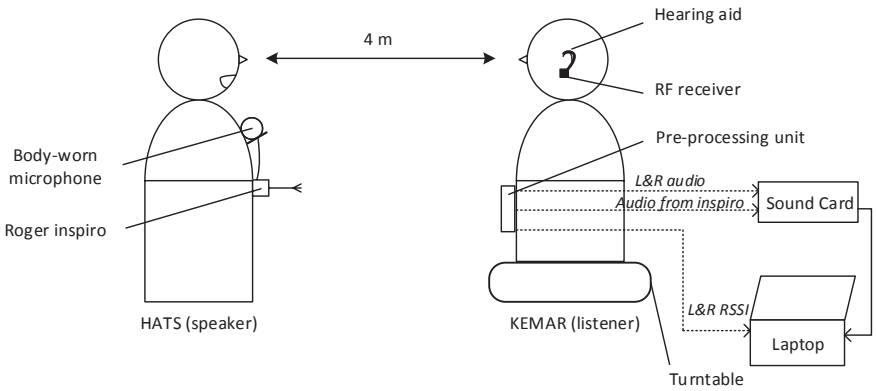


Figure 3.4 – Diagram of the acquisition setup.

in the listening room.

3.1.3 Parameter optimization

The database acquired during this experiment is used to optimize the localization algorithm. The recordings feed a Simulink model that runs the complete process offline. In this chapter, the term "simulation" refers to a run of the algorithm. Then, one has to define a method, draw the list of all the tunable parameters, select the ones with the greatest influence on the outcomes, and look for their optimal values, so as to improve the global performance of the BLA.

Parameter selection

The list of all the parameters encapsulated in the different localization blocks is the following:

- VAD block: Decision threshold,
- CE block: Number of accumulated values for the moving average,
- ILD block: Threshold of the "unknown" status & Number of accumulated values of energy before computing the ILD,
- IPD block: Distance between the 2 microphones in the mathematical model of the head (Equation 2.13) & Threshold of frame acceptance,
- RSSID block: Threshold of the "unknown" status & Leaky integrator coefficient (Equation 2.20),

- Localization block: Weightings applied on the probabilities from the IPD (Equations 2.23 and 2.24) & Number of accumulated probabilities,
- Tracking block: Values in the probability network (Equation 2.26).

This represents a number of 11 values to tune. One has to recall that the probability network is actually composed of 25 transition weightings, which gives a total of 34 values to adjust. Obviously, it is impossible to process such an amount of parameters. Practically, it has been decided to keep only 6 parameters for the tuning, and empirically fix all the other ones. These 6 selected parameters are:

1. The VAD threshold κ,
2. The number of accumulated values of energy before the derivation of the ILD γ,
3. The threshold of frame acceptance in the IPD block ξ,
4. The leaky integrator coefficient for the RSSID computation λ,
5. The number of accumulated probabilities in the Localization block ρ,
6. 2 different probability networks in the Tracking block $\mathbf{\Omega}$, one yielding a low inertia (i.e. a set of probabilities encouraging the passing from one sector to another) and one yielding a high inertia (the probabilities of staying in the current sector are increased).

One must give to each retained parameter a range of possible values. This is detailed in Table 3.3.

Block	Parameter	Min. value	Max. value	Unit
VAD	κ	25	50	Proportion of accepted frames (%)
ILD	γ	0.1	1	s
IPD	ξ	0.1	0.8	-
RSSID	λ	0.95	0.995	-
Localization	ρ	5	15	Nb of frames
Tracking	$\mathbf{\Omega}$	High inertia	Low inertia	-

Table 3.3 – List of the BLA parameters to be tuned, and their minimum and maximum values. Taken from [43].

Factorial experiment

The first step of the tuning procedure of the retained parameters is to determine and analyze their effect on the accuracy and reactivity of the BLA. The factorial experiment is a tool that enables to establish the trend and intensity of these effects for a given set of parameters [143]. With 6 variables and 2 levels for each, there are $2^6 = 64$ simulations to be conducted for each score. Figure 3.5 depicts an example of 64 observations obtained for the first set of data and a

given set of parameters. The stimuli is the female voice played in the listening room (Figure 3.5A) and in the classroom (Figure 3.5B). As expected, the listening room leads to some better couples of accuracy and reactivity than the classroom. The opposition of the 2 scores is also clearly noticeable, i.e. the more the accuracy, the less the reactivity, and conversely. Note that all the values of accuracy are always greater than 60% in both environments, highlighting the overall good performance of the algorithm. On the other hand, the reaction time is long, e.g. 2.3 s are required to get an accuracy of 80% in the listening room, and it even rises to 5.4 s in the classroom. It is not reasonable to end up with a delay of several seconds to locate the speaker. This supports the needs for an efficient optimization of the 6 retained parameters.

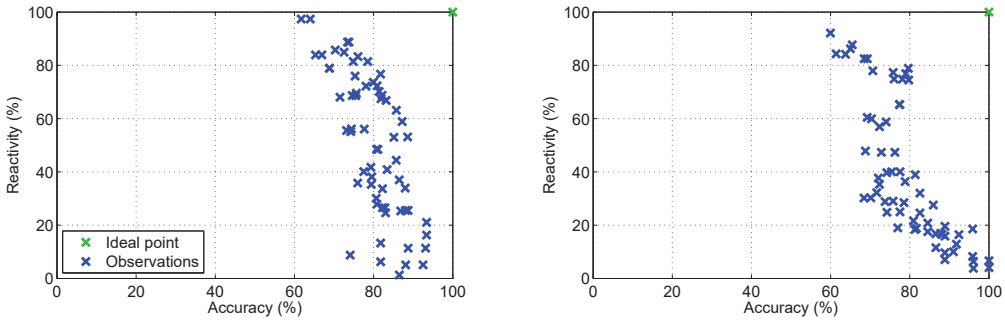


Figure 3.5 – Observation points obtained with the reported factorial experiment (blue) and the ideal point (green) in the Accuracy × Reactivity plane, for the female stimulus played in the listening room (A) and the classroom (B).

Table 3.4 and 3.5 report the numerical results from the factorial experiment, on the accuracy and reactivity respectively. They are computed for the 2 speakers' genders, in the 2 rooms, and repeated twice to get 2 different sets of measurements. The mean and SD over rooms, genders and sets of measurements are provided as well. Note that only the main effect are reported for practical reasons. The interaction effects are discussed later. The values have the same unit as the scores, i.e. percentages. An effect with a positive value means that the accuracy or reactivity augments when the parameter increases, while a negative value stands for a decrease of the score when the parameters increases. The SD gives an indication to estimate whether the average tendency is similar in both rooms, for both stimuli and for both sets.

The following comments can be made:

- All the parameters yielding an improvement of the accuracy provides a reduction of the reactivity. As expected, the diminution of κ and ξ, which translates into a smaller number of accepted frames, makes the reactivity worse but enhances the accuracy. This is because only the optimal frames for achieving the localization are kept. Some transition probabilities that favored the motion between the estimated positions dismisses the stability of the algorithm, and thus its accuracy. Conversely, some higher integration

times for the ILD (γ) and for the RSSID (λ) improve the precision of the algorithms. Also, a longer average of the sector probabilities (i.e. an increase of ρ) is beneficial for the accuracy.

- The accuracy is mainly governed by ξ, ρ and $\mathbf{\Omega}$, while the effect of κ, γ and λ, presenting an important SD relative to their value, are of less interest.
- The reactivity is also primarily monitored by ξ, ρ and $\mathbf{\Omega}$, which all exhibit some relatively low SD values. With a SD of 2.12% almost equal to its mean value (2.38%), κ does not have a reliable effect. γ and λ bring about a non-negligible influence on the reaction time, but this is of less importance compared to the 3 aforementioned parameters.

	Gender	κ		γ		ξ		λ		ρ		**Ω**	
Set		1	2	1	2	1	2	1	2	1	2	1	2
List. Room	M	-1.61	-0.99	0.02	1.52	-2.59	-1.42	0.41	1.48	1.81	2.44	-2.43	-1.53
	F	-0.45	-2.43	0.29	3.19	-1.29	-1.98	1.04	-0.22	2.63	3.13	-3.09	-3.64
Classroom	M	-3.88	-2.64	-0.01	0.47	-6.46	-3.59	0.77	2.19	7.39	6.41	-4.40	-6.25
	F	-2.19	-1.72	0.21	1.17	1.44	-2.57	-6.33	3.75	6.33	4.57	-2.31	-5.09
Mean		**−1.99**		**0.86**		**−2.78**		**1.36**		**4.34**		**−4.10**	
SD		1.06		1.09		1.66		1.21		2.14		1.74	

Table 3.4 – Results of the factorial analysis over all parameters for the accuracy. Taken from [43].

	Gender	κ		γ		ξ		λ		ρ		**Ω**	
Set		1	2	1	2	1	2	1	2	1	2	1	2
List. Room	M	4.55	4.20	-9.84	-10.8	5.78	7.60	-6.13	-3.25	-9.90	-8.15	18.40	17.20
	F	1.54	5.14	-6.71	-4.52	16.50	16.30	-6.28	-5.91	-5.76	-6.14	10.10	15.40
Classroom	M	-0.88	0.75	-1.38	-2.69	10.40	5.82	-5.38	-8.53	-10.30	-8.57	20.50	19.50
	F	1.01	2.69	-3.50	-1.34	9.52	11.00	-8.63	-4.92	-7.14	-5.77	17.10	20.40
Mean		**2.38**		**−5.10**		**10.4**		**−6.10**		**−7.72**		**17.30**	
SD		2.12		3.67		4.20		1.79		1.80		3.41	

Table 3.5 – Results of the factorial analysis over all parameters for the reactivity. Taken from [43].

The interactions between parameters were computed but are not detailed here, as they represent hundreds of combinations. It appears that the effects of κ and $\mathbf{\Omega}$ depend on each other, especially for the reactivity. This is also true for the effects of λ and $\mathbf{\Omega}$, of which the dependence is high for the accuracy in the classroom and the reactivity in the listening room.

In order to go further in the optimization of the algorithms, a new selection among the 6 parameters has to be done. Indeed, a complete optimization procedure over 6 parameters would be extremely difficult to conduct, because it would result in some models of high complexity, demanding lots of computations. It is more usual to select a small number of parameters that provide some prominent and contradictory effects on the 2 scores to optimize. The factorial experiment helps select such parameters. γ should not be kept for the analysis, as its effect is limited and highly dependent on the tested configurations. Since it appears to have more influence on the reactivity, it is chosen to fix it at its smallest level, i.e. 100

ms. κ is an unreliable parameter, in the sense that it depends on external factors on which there is no control (loudness of the speaker voice, distance to the microphone...). Therefore, it is hazardous to base the optimization procedure on a fine tuning of the VAD threshold. Fortunately, its effect is limited and it is not expected to modify the performance of the algorithm in a significant way. It is discussed in detail in part 3.2.4. Hence, κ is discarded from the analysis and set to its low level, i.e. a tolerance that reject approximately 75% of the input frames.

Although presenting an important effect on both the accuracy and reactivity, $\mathbf{\Omega}$ is neither retained for the procedure because the common methods of parameters tuning are not compatible with the management of a collection of values (i.e. a matrix), that is actually a set of numerous parameters in itself. Since κ has been fixed to its low level, which slightly favours the accuracy, it is chosen to use the probability network with the low inertia, so as to compensate and support the reactivity. Therefore, the 3 parameters kept for the next optimization procedure are ξ, λ and ρ.

Principle

The approach that consists in simulating all the possible combinations of the 3 retained parameters and deriving the corresponding distance D (Equation 3.4) is not satisfying. In fact, a simulation takes several minutes to be performed and it would result in an excessive amount of time. Therefore, smarter methods have been implemented so as to converge to an optimal point with coordinates $(\xi_{\text{OPT}}, \lambda_{\text{OPT}}, \rho_{\text{OPT}})$ in a more efficient and rapid way.

The objective is to estimate an unknown function g so that:

$$D = g(\xi, \lambda, \rho). \tag{3.5}$$

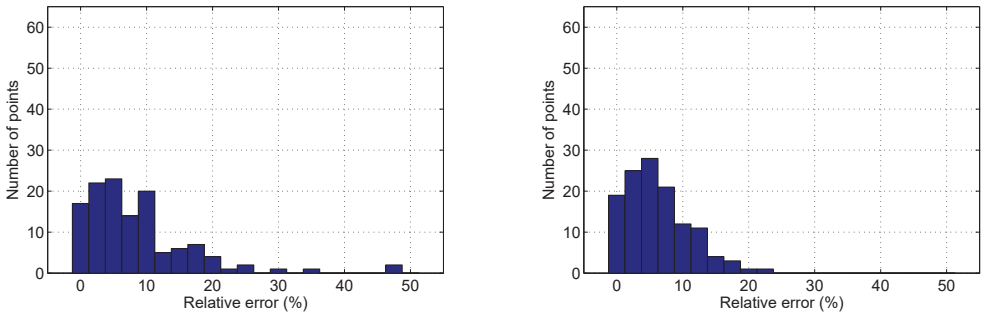
Note that one function g must be considered in the different test configurations (both genders in both rooms).

The ultimate goal is to come up with one point of which the coordinates are $(\xi_{\text{OPT}}, \lambda_{\text{OPT}}, \rho_{\text{OPT}})$, corresponding to a local minimum of g, that is:

$$(\xi_{\text{OPT}}, \lambda_{\text{OPT}}, \rho_{\text{OPT}}) = \underset{(\xi,\lambda,\rho)}{\arg\min}\, g. \tag{3.6}$$

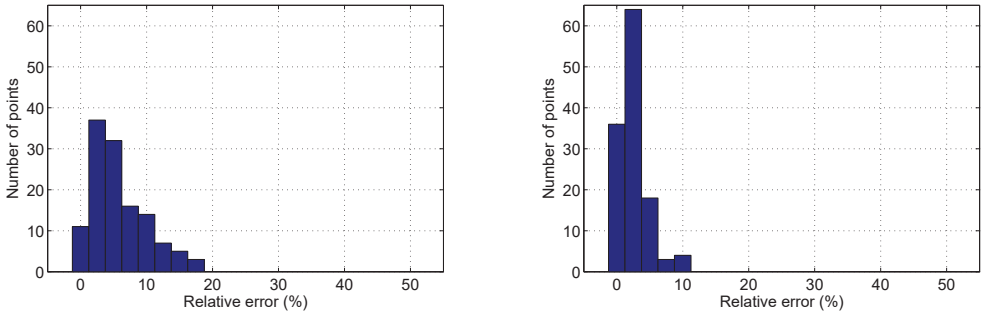
In the following, the 3 parameters are varying from -1 to 1, delimiting a linear scale of which the boundaries are the minimum and maximum values reported in table 3.3.

Response surface design



(a) (b)

Figure 3.6 – Distribution of the relative error between the real distance and the quartic model for 125 observations in the classroom, with the male (A) and female (B) stimuli. Taken from [43].



(a) (b)

Figure 3.7 – Distribution of the relative error between the real distance and the quartic model for 125 observations in the listening room, with the male (A) and female (B) stimuli. Taken from [43].

The *response surface design* [120, 214] suggests to model the function to optimize using an approximated known non-linear function. It facilitates the search of the local minima in a faster way. To this end, some reference data have been collected. They correspond to the accuracy and reactivity scores from 125 combinations of the 3 parameters, regularly spaced between the minimum and maximum value of each parameter (5 levels, i.e. $\xi = [-1 \quad -0.5 \quad 0 \quad 0.5 \quad 1]$, same for λ and ρ). These scores serve as a reference dataset to estimate the non-linear function. 4 sets of 125 observations have been computed, for the following configurations: male speech in the classroom, female speech in the classroom, male speech in the listening room, female speech in the listening room. A 4^{th} order polynomial regression has

been chosen for the function g, and the 35 resulting coefficients a were estimated such that:

$$g = a_0 + a_1\xi + a_2\lambda + a_3\rho + a_4\xi\lambda + a_5\xi\rho + ... + a_{33}\xi^4 + a_{34}\lambda^4 + a_{35}\rho^4 + r, \tag{3.7}$$

where r denotes the residues resulting from the modeling error.

Configuration	Male speech in classroom	Female speech in classroom	Male speech in listening room	Female speech in listening room
Mean relative modeling error (%)	8.47	6.26	5.81	2.53

Table 3.6 – The mean related modeling error in the 4 reported configurations. Taken from [43].

An order of 4 for the polynomial regression has been fixed because it leads to the smallest errors between the real and approximated distances. The distribution of the relative error (expressed as a percentage) among the 125 references of each room is shown on the bar graphs of Figure 3.6 (classroom) and Figure 3.7 (listening room). In all cases, the majority of the points shows some error values smaller than 10%. The distribution is more spread in the classroom, i.e. larger errors are reached. The more complex acoustic conditions in this environment make it more difficult to model the distance D with a polynomial. There is no noticeable difference in the quality of the modelling between the male and female stimuli. In the listening room, the approximation of the female stimulus with a quartic function is better than with the male speech.

Table 3.6 gathers the mean relative errors obtained in the 4 different configurations over the 4 reference datasets of 125 points. The values confirm the results previously reported. In particular, the female speech in the listening room appears to be significantly well modeled by the polynomial.

Thanks to the polynomial model, it is possible to estimate the distance D for any coefficient combination (ξ, λ, ρ), without requiring the time of a complete simulation. Figure 3.8 and 3.9 depict the estimated Euclidian distance D as a function of ξ and ρ for the minimum and maximum levels of λ, in the classroom (Figure 3.8) and in the listening room (Figure 3.9), with the male and female stimuli. A step of 0.01 is taken for the variation of the 2 parameters ξ and ρ, in order to derive the quartic function on a sufficiently wide mesh. In both environments, the female speech provides the highest error. A remarkable observation is the fact that λ has a strong effect in the classroom, i.e. the rise of λ decreases the performance of the algorithm. Therefore, the leaky integrator coefficient has to stay low. That is equivalent to say that it is not worthy to accumulate too much previous RSSID values to calculate the current one, taking into account both the accuracy and reactivity of the BLA. The absence of influence in the listening room is most probably due to the fact that the quality of the RSSID is such that the leaky integrator coefficient does not matter. Intermediate values of λ have been tested to confirm this preservation. As the trend is the same, the value of λ is fixed to its small value in

the following, i.e. $\lambda_{OPT} = 0.95$.

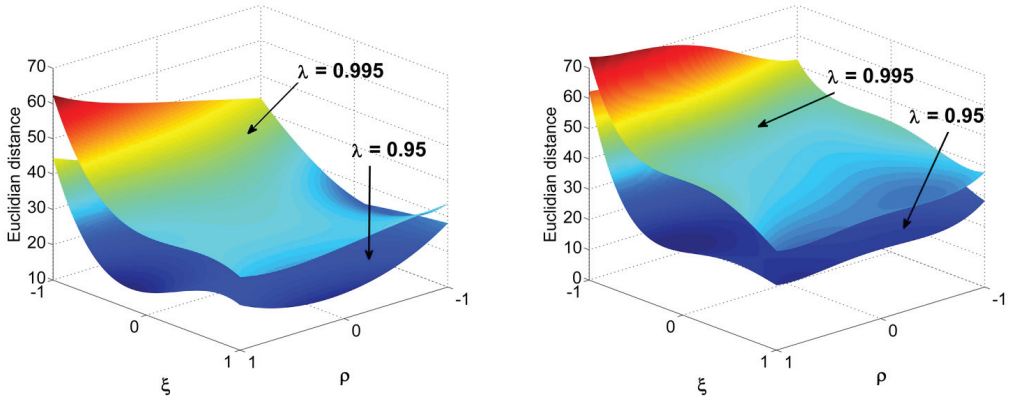


Figure 3.8 – Evaluation of the model for various values of ξ and ρ (steps of 0.01) and for the minimum and maximum levels of λ, in the classroom, with the male (A) and the female (B) stimuli. Taken from [43].

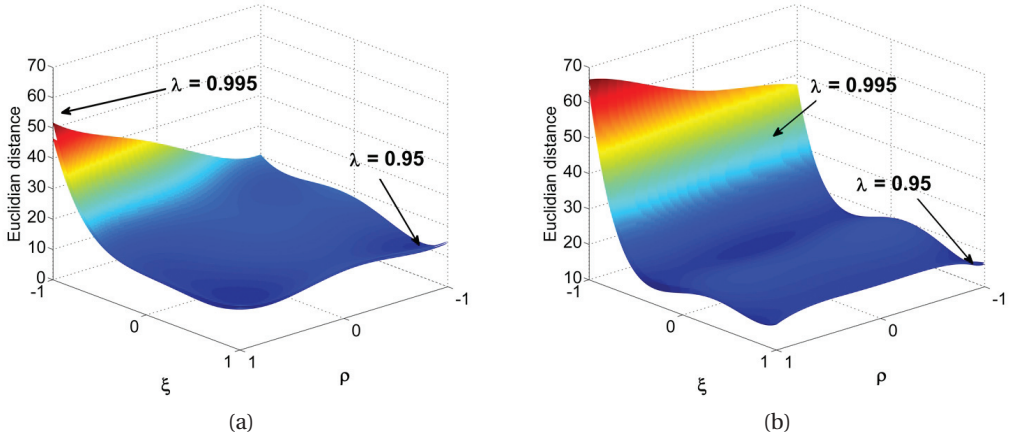
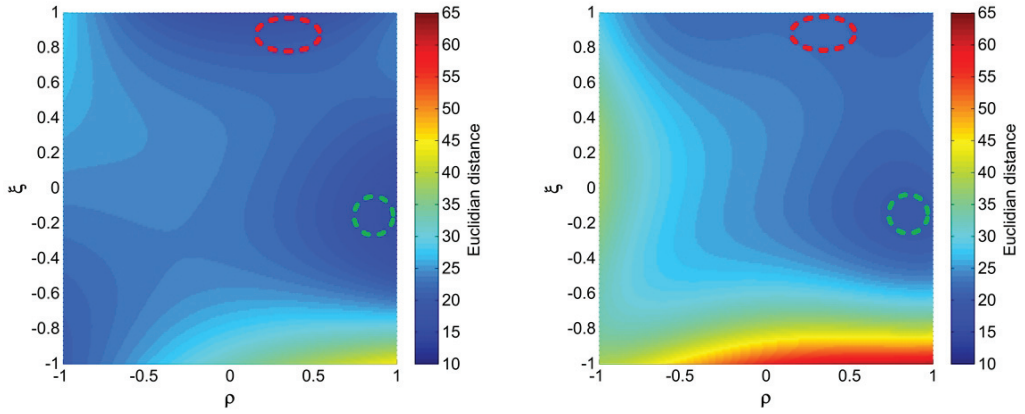
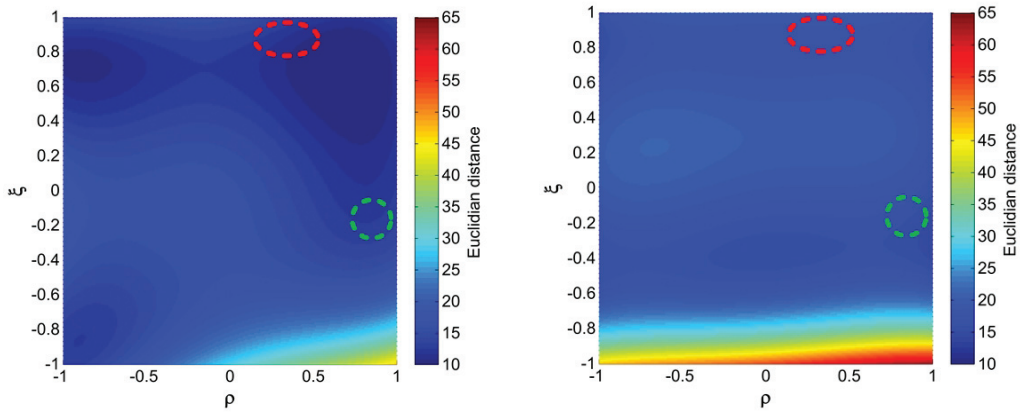


Figure 3.9 – Evaluation of the model for various values of ξ and ρ (steps of 0.01) and for the minimum and maximum levels of λ, in the listening room, with the male (A) and the female (B) stimuli. Taken from [43].



(a) (b)

Figure 3.10 – Computation of the model for various values of ξ and ρ (steps of 0.01), in the classroom, with the male (A) and the female (B) stimuli. The red and green circles highlight some optimal areas common to the 4 configurations. Taken from [43].



(a) (b)

Figure 3.11 – Computation of the model for various values of ξ and ρ (steps of 0.01), in the listening room, with the male (A) and the female (B) stimuli. The red and green circles highlight some optimal areas common to the 4 environments. Taken from [43].

Figure 3.10 (classroom environment) and Figure 3.11 (listening room environment) show the map of the values of the polynomial as a function of the coefficients ξ and ρ. 2 areas of interest have been highlighted with the red and green circles, because they provide small values of the distance in the 4 configurations. Although the green area seems to be the most promising one, it actually appears to be particularly affected by the modeling error, after some simulations with the acquired data. That is, the Euclidian distance D is underestimated in this region. On the contrary, the red area provides really interesting outcomes, since the distance has been

sometime overestimated by the model. A finer exploration has been conducted in this region so as to determine the optimal combination of the 2 parameters ξ and ρ.

Block	Parameter	Optimal value	Unit
VAD	VAD threshold κ	25	Nb of passing frames (%)
ILD	Integration time γ	0.1	s
IPD	Acceptance threshold ξ	0.765	-
RSSID	Leaky integrator coefficient λ	0.95	-
Localization	Frame accumulation ρ	12	Nb of frames
Tracking	Probability network $\mathbf{\Omega}$	Low inertia	-

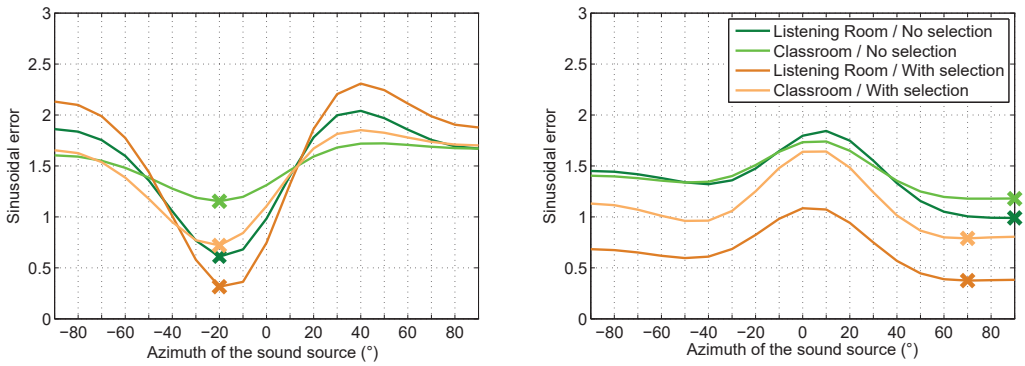
Table 3.7 – Optimal values of the BLA parameters.

Table 3.7 summarizes the values taken by the different parameters after the optimization. It is now relevant to compute the final performance of the BLA.

3.2 Post-optimization performance

After having conducted the optimization of the parameters, the performance of the resulting algorithm is assessed. Several factors are investigated: the process of IPD selection, the fusion of the different localization cues, the accuracy and reaction time, the VAD effect, and the influence of the size of the head.

3.2.1 Interaural phase difference



(a) (b)

Figure 3.12 – Average sinusoidal error, with the speaker located at -20° (A) or at 70° (B). The results in the listening room are depicted in dark colors, while the ones measured in the classroom are in light colors. The orange lines represent the case when the frame selection is applied and the green lines correspond to the case where all frames are processed. The crosses highlight the smallest error, i.e. the most likely DOA. Taken from [43].

Figure 3.12 shows the average sinusoidal error, as computed with Equation 2.16, when the sound source is located at -20° (Figure 3.12A) or at 70° (Figure 3.12B) for the various tested azimuths in the FHP. The results in the listening room are depicted in dark colors, while the ones measured in the classroom are in light colors. The orange lines represent the case when the frame selection is applied (Equation 2.17) and the green lines correspond to the case where no selection is applied. The frame selection succeeds in rejecting a majority of frames for which the IPD pattern is far away from the model and would lead to wrong localization. The range of error between the higher and lower probable azimuths is enlarged, which helps identify the current location of the source. This also limits the risk that a wrong speaker's position is determined under a fixed-point resolution (truncation error). The crosses show the smallest errors, which represent the IPD-based determined locations. When the source is at -20°, the frame selection is not necessary to come up with the correct localization, but the minimum error is made more visible. However, when the speaker is located at 70° , the process of frame selection avoids a wrong localization at 90° in both environments.

Three one-way between-subjects ANOVAs have been computed on the mean over 45600 values of the sine error (Equation 2.16) associated with the most likely DOA, when the algorithm is run offline on the acquired data. The goal is to analyze the effect of the 3 factors (azimuth, gender and room), in order to reveal which ones have a significant influence on the IPD-based localization. The results of the ANOVAs are reported in Table 3.8. As a great number of outputs can be derived, the choice of a type-I error of 1% ($\alpha = 0.01$) seems to be reasonable. A Brown-Forsythe correction has been applied, since the data did not fulfill the homogeneous variance assumption. One can notice that the precision of the IPD-based localization depends on the 3 factors: azimuth ($F_{18,45481.7} = 51.187, p < 0.01$), gender ($F_{1,45593.8} = 8.726, p < 0.01$), and room ($F_{1,45597} = 335.216, p < 0.01$). The results appear to be better for the frontal azimuths rather than for the lateral ones, which is discussed in the following. There is also a significant effect of the room, which is not a surprising outcome as the listening room provides some better acoustic conditions. Finally, the statistical difference between the male and female conditions can be considered as a weak point, since it reveals that the IPD-based localization depends on the speaker's pitch.

Factor	d.f. 1	d.f. 2	F	Prob>F
Azimuth	18	45481.7	51.187	<0.001
Gender	1	45593.8	8.726	0.003
Room	1	45597.0	335.216	<0.001

Table 3.8 – Results of the 3 one-way ANOVAs, showing the effect of the 3 factors (azimuth, gender and room) on the IPD-based localization. The significant effects are in red ($\alpha = 0.01$).

3.2.2 Multimodal localization

In addition to the IPD, the BLA utilizes the ILD and RSSID to localize the speaker. Figure 3.13 shows the average probabilities of being in one of the 5 spatial sectors for the different

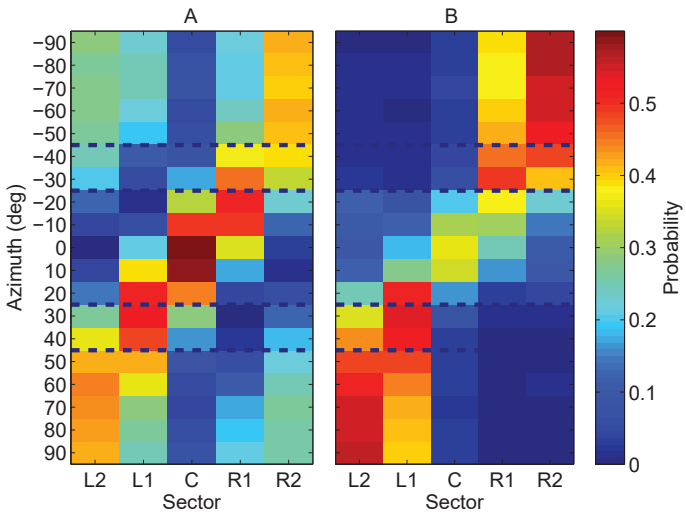


Figure 3.13 – Probabilities of being in one of the 5 spatial sectors, for all tested speaker's positions (male speech) between -90° and 90° in the listening room. It shows the average probabilities over all analysis frames using only the IPD cues (A), and the additional contributions of the ILD and RSSID (B). The dotted black lines represent the sector boundaries.

speakers' positions between -90° and 90°. The corresponding test condition is the male speech stimulus played in the listening room. Figure 3.13A depicts the average IPD-based probabilities, computed according to Equation 2.22. The dotted lines represent the sector boundaries defined in the BLA. Figure 3.13B displays the weighted scores including the contribution from the ILD and RSSID (Equation 2.23). Note that the values $S_W(\Theta_i)$ have been normalized for all frames so that they sum up to 1, for comparison between the 2 panels of the figure. Without the ILD and RSSID contributions, the algorithm is capable of matching the speaker's location with the adequate spatial sector, although some mistakes may occur. After the ILD/RSSID weighting operation, the risk of a serious localization error (e.g. localizing in the extreme left sector while the speaker is located in the extreme right one) is dramatically reduced. The downside is that improper weightings are occasionally given to the lateral sectors when the source is actually located in the front. Nevertheless, the probabilities remain sufficiently high to accurately localize the speaker.

3.2.3 Accuracy and reaction time

The performance of the BLA with the final combination of its parameters has been evaluated in terms of accuracy and reactivity. Figure 3.14 depicts the accuracy and reactivity scores of the algorithm in the classroom, for the male (blue) and female (pink) speech, according to the selected configurations and averages defined in Tables 3.1 and 3.2.

Concerning the accuracy (Figure 3.14A), the results appear to be well above the chance level (20%) for each spatial sector. As seen in the previous part, the performance depends on

the azimuths. Here, the lowest scores are located in the central spatial sector, which can be explained by the absence of contribution of the ILD and RSSID cues in this spatial area. The sectors on the left appear to provide better scores than the ones on the right. This is most likely due to the fact that the KEMAR was closer to the wall on its right than to the windows on its left. Therefore, more reflections may have occurred on the right than on the left side. The average accuracy in the listening room is equal to 89% for the male speech and 84% for the female stimulus.

Three one-way between-subjects ANOVAs have been conducted so as to assess the effects of the sector, gender of the speaker and room on the accuracy scores. The results are reported in Table 3.9. The ANOVAs fail to show any significant influence of these 3 factors: sector ($F_{3,18} = 3.164$, $p = 0.045$), gender ($F_{1,18} = 2.495$, $p = 0.132$), and room ($F_{1,18} = 3.251$, $p = 0.088$). One can conclude that the significant effects that have been observed on the IPD-based localization are lost once the entire localization algorithm is applied. This suggests that the algorithm is somehow less sensitive to these external factors, probably because of both the lowest resolution (5 sectors against 10° steps) and the multi-modal approach.

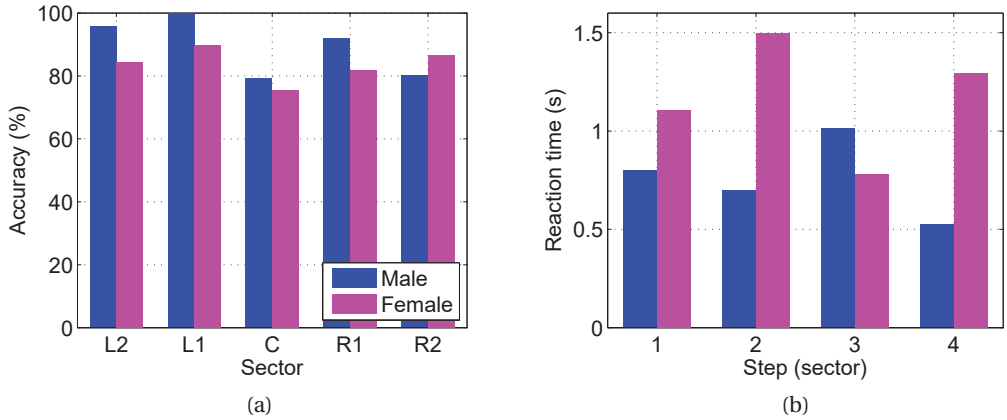


Figure 3.14 – The accuracy of the optimized BLA (A) in the classroom in each spatial sector, for the male speech (blue) and female speech (pink), and the reactivity (B) for the different tested steps (in number of sectors).

Figure 3.14B depicts the reactivity scores of the BLA. The outcomes are reported in Table 3.10. The results are given for different step sizes (given in number of sectors). The mean reaction time in the classroom is 0.76 s for the male speech and 1.17 s for the female speech. Three one-way between-subjects ANOVAs have been computed to evaluate the effect of the step size, gender and room. This analysis fails to show any statistical effect of the gender ($F_{1,14} = 2.236$, $p = 0.157$), the room ($F_{1,14} = 4.360$, $p = 0.056$) and the step size ($F_{3,12} = 0.928$, $p = 0.457$) factors, which is similar to what was found with the accuracy.

Factor	d.f. 1	d.f. 2	F	Prob>F
Sector	4	18	3.164	0.045
Gender	1	18	2.495	0.132
Room	1	18	3.251	0.088

Table 3.9 – Results of the 3 one-way ANOVA, showing the effect of the 3 factors (sector, gender and room) on the accuracy. There is no significant effect ($\alpha = 0.01$).

Factor	d.f. 1	d.f. 2	F	Prob>F
Step	3	12	8.418	0.003
Gender	1	14	1.825	0.198
Room	1	14	0.777	0.393

Table 3.10 – Results of the 3 one-way ANOVA, showing the effect of the 3 factors (sector, gender and room) on the reaction time. The significant effects are in red ($\alpha = 0.01$).

3.2.4 VAD effect

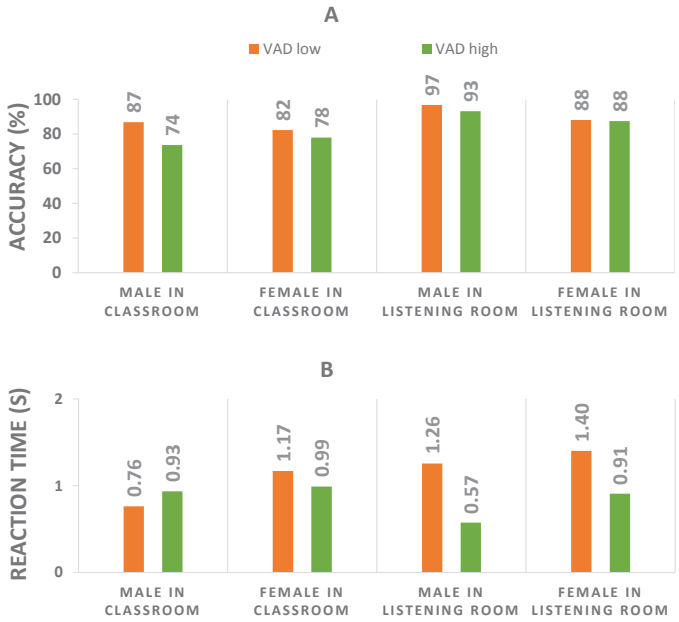


Figure 3.15 – Accuracy (A) and reactivity (B) performance of the BLA depending on the VAD threshold, $\kappa = 25\%$ in orange and $\kappa = 50\%$ in green.

The optimization procedure has been performed with the VAD threshold κ at its low level (see part 3.1.3). However, κ depends on external factors, such as the distance from the mouth to the microphone, or the SPL of the speaker's voice. A brief analysis studied the effect of switching κ to its high level, so as to approximate the condition of a closer microphone or a

louder voice. Note that this simulation is not completely accurate, since it does not improve the SNR, as it would be the case if the speaker would speak louder or closer to the microphone. The results of this study are shown on Figure 3.15A for the accuracy, and on Figure 3.15B for the reactivity.

Overall, increasing κ yields a reduction of the accuracy, with a more pronounced effect in the classroom than in the listening room. Conversely, the reaction time is reduced with the rise of κ, except for the male speech in the classroom. This is especially true in the listening room. These results are consistent with what would be expected. Performing the localization on a higher number of frames implies to consider potentially adverse segments (see Chapter 2.4.1). Because the acoustic conditions are worse in the classroom, the deterioration is more pronounced. Since the reactivity score include the notion of accuracy, the changes are not remarkable in the classroom. On the contrary, a higher value of κ would be recommended in the listening room, because of the better acoustical conditions. Considering the fact that the targeted environments are closer to the classroom characteristics, one prefers to keep the best parameters in this situation, i.e. $\kappa = 25\%$.

3.2.5 Head-size effect

Figure 3.16 – 3D printed heads used to study the effect of the head size on the algorithm performance. The left head is referred as SMALL (80% of the original size), the one in the center is the MIDDLE (100% of the original size) and the right one is called BIG (120% of the original size).

Since the BLA is based on a model parameterized by a fixed average inter-ear distance (Equation 2.13), it is prominent to test the robustness of the algorithm against different head sizes. In particular, the IPD observed with a smaller head have lower values and would lead to excessively centered localization. To this end, three 3D printed heads have been designed and used for the algorithm evaluation, as shown on Figure 3.16. They are referred as SMALL (80% of the original size), MIDDLE (100% of the original size) and BIG (120% of the original size) in the following. Having the same size as the original KEMAR head, the MIDDLE head has been used to study the effect of resorting to 3D printed heads on the algorithm performance, and to ensure that this approach is legitimate. Then, the SMALL and BIG heads allow to

investigate the effect of the head size on the localization accuracy. The 3 heads were filled with electromagnetic absorbers.

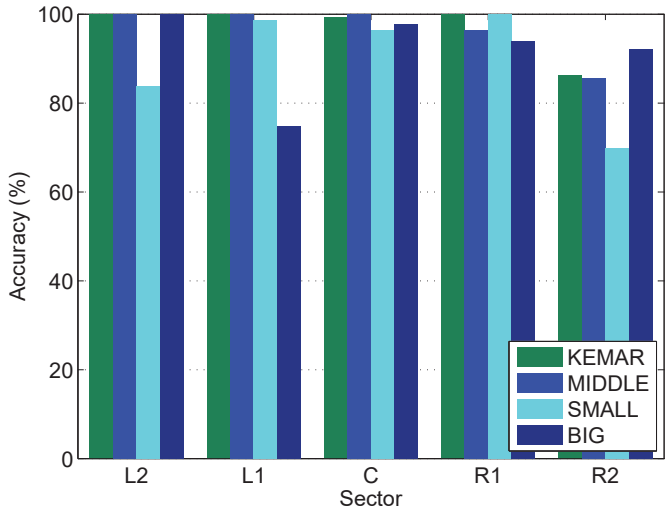


Figure 3.17 – Accuracy scores of the BLA in the listening room with the male speech for different head types and sizes (KEMAR in green, and 3-size printed head in different tint of blue).

The investigation of the head size effect on the accuracy has been performed in the classroom only, with both stimuli. Figure 3.17 shows the accuracy scores obtained in this configuration with the original KEMAR head, and the MIDDLE, SMALL and BIG 3D printed heads. The male and female stimuli have been used, and the mean of both is depicted here. A comparison between the results from the KEMAR and MIDDLE head suggests that the use of a 3D printed head is relevant in this experiment. A one-tailed Dunnett test, taking the KEMAR as the reference, has been performed to test the null hypothesis H_0 that there is no difference of accuracy among the different heads, against the alternative hypothesis H_1 that the size of the head has a detrimental effect on the performance of the BLA. The test failed to show any significant effect (KEMAR vs MIDDLE: $p = 0.704$, KEMAR vs SMALL: $p = 0.238$, KEMAR vs BIG: $p = 0.631$), so the alternative hypothesis is rejected.

3.2.6 Pre- and post-optimization performance

The primary limitation of the optimization procedure is that it has been based on a certain set of data, for some specific environments and stimuli. Indeed, it is neither reasonable nor possible to acquire data for a large number of configurations. However, the recordings of several databases (2 in the listening room and 3 in the classroom) for the same setup allows a partial assessments of the generalization of the BLA final performance. Moreover, it is possible to assess the efficiency of the optimization by comparing the outcomes of the algorithm before and after the procedure. The results of the pre-optimized algorithm corresponds to an empirically-found combination of parameters, given in Table 3.11.

Parameter	Pre-optimization value	Post-optimization value
κ (%)	25	25
γ (s)	0.1	0.1
ξ	0.8	0.765
λ	0.95	0.95
ρ (Nb. of frames)	12	15
$\mathbf{\Omega}$	Low inertia	Low inertia

Table 3.11 – Comparison of the BLA parameters before and after the optimization procedure.

Figure 3.18 shows the accuracy scores (Figure 3.18A) and reaction times (Figure 3.18B) for the 3 sets of data in the classroom, before and after the optimization. The outcomes are averaged over the male and female stimuli. A precise accuracy is ensured for the 3 sets of data, with scores greater than 82%. On the other hand, the reactivity of the BLA significantly varies between the different datasets. For instance, there is a 1.5-s gap between the reaction time in the first and in the second database. These observations indicate that the accuracy is robust against the 3 sets of data, while the reactivity highly depends on them. It is problematic since it means that the performance of the algorithm cannot be generalized, even in the same environment.

The optimization procedure does not manage to solve this issue, but does not worsen it either. One can notice that the optimization has not improved the performance of the BLA in a substantial way, except for the reaction time that is reduced by 160 ms in the first set of data. This indicates that the empirically-found set of parameters already provided good performance before optimization.

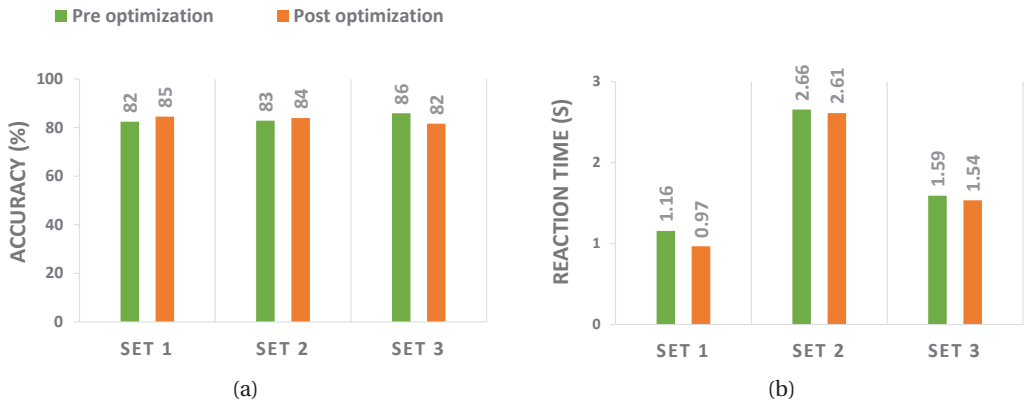


Figure 3.18 – The accuracy (A) and reaction time (B) of the BLA before (green) and after (orange) the optimization of the BLA in the classroom, for the 3 sets of data. The results are averaged over the male and female stimuli.

The BLA with its optimized combination of parameters can now be implemented on the targeted hardware, as detailed in Appendix D.

3.3 Conclusion

This chapter has covered the optimization of the BLA. In fact, the localization algorithm provides a large number of parameters that can be tuned to come up with the best possible performance. The successive actions required to pursue this work have been:

- The definition of the scores to maximize. Within the frameworks of the localization algorithms, the 2 main features to optimize are the accuracy of the localization (to be maximized) and the time needed to react to a motion of the source (to be minimized). Both scores are antagonist. Indeed, a good accuracy usually requires a long time to output the right position. On the contrary, the speeding up of the computation increases the risk of instability of the results. This shows the importance to find an accurate trade-off between both of them, to allow a correct localization in a reasonable amount of time. The accuracy and reactivity of the BLA have been defined and used to obtain a good balance between the 2,
- The acquisition of a database of signals in a classroom and in a listening room, intended to optimize and assess the BLA with “real-world” signals,
- The identification and selection of the parameters one should focus on, thanks to the outcomes resulting from a factorial design of experiments,
- The implementation of suitable and efficient methods of optimization to end up with the best performance. This includes a simplex optimization, then a response surface design to tackle the problem of the local minima.

The major limitation of this optimization work is that it has been performed on a precise database only, since it was unrealistic to acquire some signals in various acoustic conditions. This poses the question of the generalization of the performance obtained with the optimized set of parameters, which actually appears to depend on the different sets of measurements. Anyway, the knowledge of the final performance of the BLA in the chosen environments is necessary to characterize the behavior of the algorithm. The effect of several contributions and factors have been assessed to estimate and justify the decision taken during the research and optimization of the BLA. This includes the influence of:

- The process of frame selection at the end of the IPD block. It prevents the algorithm from making some false localizations in certain cases. The main associated drawback is the added latency,
- The contribution of the ILD and RSSID, which considerably decreases the risk of ending with a serious localization error (e.g. extreme left instead of extreme right),
- The VAD selection that must not be too much tolerant. Indeed, a permissive acceptance of potentially harmful speech frames degrades the localization precision, while not improving the reaction time,

- The head size effect, motivated by the fact that the BLA rests upon a spherical model of the head, with a fixed head radius that may not be convenient for little or big head sizes. Thanks to the use of 3D printed heads, it has been possible to show that the algorithm is robust against this factor.

Overall, the statistic tests that have been performed on the accuracy and reaction time of the BLA failed to show any effect of the gender, room, and head size, somehow suggesting that the algorithm is robust against these 3 factors, although this cannot be fairly concluded from the results. In a typical classroom, the BLA is able to accurately locate the speaker around 85% of the time and requires between 1 and 2.5 s to output this accurate localization. One has to keep in mind that this information is obtained with a certain experimental configuration. Especially, the speaker-to-listener distance equalled 4 m, and it is expected that the algorithm can reach quite a bit better performance with shorter distances.

4 Development of a binaural spatialization algorithm

This chapter introduces the idea of applying some techniques of binaural spatialization for HI aided subjects. The developed *binaural spatialization algorithm* (BSA) reported here takes as input the estimated location of the speaker and the current resolution of the BLA. Then, the spatial rendering of the clean speech signal coming from the body-worn microphone is performed. The final objective is to allow the impaired listeners to localize the talker, and match the acoustic stimulus with the visual one. There are 2 main topics of research that are investigated in this chapter. The first one is the determination of a simple and efficient approach for spatialization, which respects the constraints demanded by the hardware. The second is related to the precaution that one must take when rendering a spatial effect in the HAs of a HI listener.

Part 4.1 reports the applications and main existing methods for achieving various qualities of binaural spatialization. Part 4.2 details the successive techniques one should apply when designing some low-cost spatial filters. The major original contribution of this chapter is the object of part 4.3, and is related to the definition and subjective evaluation of the limitation of the HRTF magnitude, of which the aim is to take into consideration the targeted hardware and end-users. Part 4.4 deals with the choices made for the spatial filter design, implementation and interpolation, in order to guarantee a real-time rendering. The final BSA has been implemented on the embedded prototype, so that some informal listening tests could be conducted to get the first impression of NH users. This is reported in part 4.5. The conclusions of the chapter are drawn in part 4.6.

4.1 Introduction

Here are introduced the most current applications of the binaural spatialization methods. Different qualities of spatial rendering are then described and discussed.

4.1.1 Applications

Appendix A.2 highlights the importance of spatial hearing for speech intelligibility and localization. Sound *spatialization*, also known as auralization, denotes the process consisting in rendering a certain recorded or generated audio stimulus as if it has been originated from a real sound source located somewhere in space. The objective is to artificially recreate a natural spatial hearing. One distinguishes 2 families of spatialization techniques, which are through loudspeaker reproduction and headphone rendering [151]. The first refers to the multichannel playback of a recorded or synthesized sound field over a certain number (at least 2) of loudspeakers. The most famous techniques are the stereophony, the ambisonics and the wave field synthesis [207, Chap. 4]. Such approaches are dedicated to large audiences. The second approach for 3D sound reproduction uses the rendering through headphones. It is called *binaural spatialization*. The concept of binaural spatialization appeared in the seventies, when Plenge [189] suggested to reproduce some binaural recordings of sound signals with earphones. 2 years later, Platte and Lawis [188] came up with the notion of HRTF-based binaural spatialization, i.e. the use of spatial filters based on the listeners' HRTFs.

Applications of spatialization are plentiful. With the aim of improving the intelligibility and segregation of multiple sound signals, Begault and Wenzel [17] provide spatialization in the headphones of a pilot in a cockpit. Thanks to a head tracker, the spatialization process can adapt dynamically. However, Freeland *et al.* [74] recall that the underlying processing has to be computationally simple to ensure a real-time procedure. No mention of a localization process preceding the spatialization stage is mentioned in the literature, to the knowledge of the author, and head tracking seems to be the only reported way of achieving dynamic spatialization. Other contexts of spatial audio applications are the entertainment industry (cinema, home theater, immersive video games...) [74] and the human-machine interactions (virtual reality, tools for helping visually impaired subjects...) [151]. Very recently, Lopez *et al.* [141] has developed an application for smartphones and tablets for videoconferencing. It allows a participant to define a spatial location for each speaker in a virtual meeting room via a touch screen. Their corresponding voices are then spatialized in the chosen direction. This is to enhance the intelligibility and localization of the different speakers involved in the videoconference.

4.1.2 Lateralization and spatialization

When it comes to spatialization, one must make the difference between the 1D spatialization (*lateralization*), 2D spatialization (*decorrelation*) and 3D spatialization, which is the concept developed in this thesis. Spatial location of virtual sound sources can be simulated applying a pair of ITD and ILD. This is the principle of lateralization, which creates some sound images situated along an imaginary horizontal line that links the 2 ears in the head. This phenomenon is known as *internalization* [172, Chap. 7]. The actual location of the sound image is controlled by the amount of ITD and ILD that are digitally introduced. Blauert [22, Chap. 2.4] explains

that "*the auditory event migrates toward the ear at which the signals appears first*". He also indicates that the virtual source moves "*towards the ear to which the stronger signal is being present*". Figure 4.1A illustrates the perceptual effect of lateralization. The position of the 3 sound events are controlled by the amount of ITD/ILD. Note that there exist some upper threshold values of ITD and ILD above which the auditory event remains at the leading and/or louder ear. Lateralization is a poor way of achieving spatial hearing because it restricts the auditory image to only one dimension. Moreover, as it is not possible to push sounds outside of the head, lateralization does not enable to recreate a credible sound perception, as it is experienced in the daily life. The spectral cues are actually missing in this process.

It is possible to go from 1D to 2D lateralization by decreasing the IC, which expands the sound image up to the entire volume of the upper head [170, Chap. 10]. The control of the IC monitors the width of the auditory object [69]. Indeed, the sound images are perceived wider and spread out with the decrement of this cue (Figure 4.1B). A commonly used technique for achieving decorrelation is the combined use of phase inversion and multiple delay lines [16]. Note that the decorrelation process has almost no effect on the perceived location of the sound source, which remains monitored by the amount of ITD and ILD. Although decorrelation provides improvements and increases the spatial rendering obtained with lateralization, an unavoidable internalized perception remains.

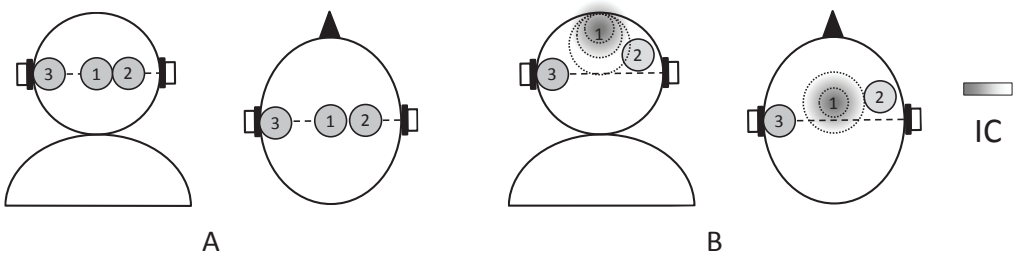


Figure 4.1 – Principle of lateralization (A) and decorrelation (B) processing of 3 virtual sound sources. Taken from [39], and inspired by [69, page 47].

The 3D binaural spatialization is a well-established technique that aims to reproduce through headphones the acoustic sound field generated by sound sources in space at both ears. The principle is to apply digital filters to the signals of the left and right channels. The filters model the HRTFs (see Appendix A.2.1) of a certain listener corresponding to the actual relative location of the sound source [16].

Wightman and Kistler [246] describe a precise and detailed process for measuring HRTFs, which show that the measurement on a subject is a long and tedious procedure. That is why it is usual to utilize generic HRTFs, measured on a manikin or a so-called "good localizer" subject [17]. In fact, Wightman and Kistler [247] report some localization test results supporting the idea that there naturally exist some "good" and "bad" localizers. Through HRTF comparisons, they show that the "bad localizers" are actually the ones whose HRTFs are the smoothest. This

encourages the selection of HRTFs from the “good localizers" for impersonalized spatialization processing (i.e. filtering signals with HRTFs different from those of the listener). Wenzel *et al.* [243] conduct some localization tests showing that individual listeners do not need to hear with their own HRTFs to get some accurate spatial information in the horizontal plane, in spite of a pronounced inter-subject variability. In particular, they show that the “good localizers" perform almost similarly with both real and spatialized sound localization as soon as the HRTFs used are the ones of a “good localizer". The “bad localizers" show poor localization performance whatever the HRTF origin. No significant effect of the use of non-individualized HRTFs on the localization in azimuth and *externalization* (i.e. perception of the sound image outside of the head) is found by Begault *et al.* [14] either. This also holds in the study led by Drullman and Bronkhorst [61], who report no difference on the localization performance of 12 subjects in the FHP. The bandwidth of the stimuli was reduced to 4 kHz, which bypasses a great deal of the spectral cues. Overall, all the aforementioned authors report that generic HRTFs do not degrade the localization accuracy in the FHP but that it leads to bad localization performance when considering elevation and front/back confusion. This is a promising information for this research, since only the FHP is considered.

Hartmann and Wittenberg [93] study in detail the link between externalization and different spatial cues, such as the IPD, ITD and ILD. This has been achieved with subjective tests on 4 listeners. They were asked to distinguish between real and virtual sound sources. The stimuli were processed so as to present some IPD, ITD or ILD at various frequencies. One of the major results of this study is that it has been possible to control the distance of the virtual source varying the frequency above which no IPD is applied. That is, the lower the frequency above which IPD is set to 0, the more the sound source is perceived close to the head, and even inside the head. This confirms the hypothesis of Blauert [22, Chap. 2] that internalized localization is a part of a continuum from externalized and distant images to in-the-head located images. However, the results must be taken with care because of the limited number of subjects involved in the experiment.

A recent investigation provides interesting outcomes and conclusions in the field of spatialization based on non-individualized HRTFs. Mendonça *et al.* [167] focus on the possibility to resort to training periods in order to make listeners more familiar with generic HRTFs. They perform localization test through headphones with 4 inexperienced subjects, using samples of pink noise spatialized in different azimuths, with the HRTFs from a KEMAR. From this experiment, they conclude that a simple exposure is not sufficient for significantly enhancing the localization ability, which confirms the results previously by Hofman *et al.* [97]. The global error averaged over all listeners and azimuths falls from 15.67° down to 8.44° after a training period of about 20 minutes. This reveals that significant improvements can be achieved with short periods of training, thanks to the brain plasticity to take over acoustic features from another subject. The authors conclude that “*virtual sounds processed through non-individualized HRTFs should only be used after learning sessions*”. Again, only 4 subjects were tested and the results should not be generalized.

4.2 Methods for the implementation of spatialization

In this section, various interesting techniques for achieving binaural spatialization are considered. Indeed, their applications may be relevant in the design of the algorithm. The progressive study of these techniques is detailed on Figure 4.2. The minimum-phase property of HRTFs is discussed in part 4.2.1, as a tool to help the design of spatial filters, which is the object of part 4.2.2. The principle of *frequency warping* is introduced in part 4.2.3. It allows to take into account the frequency resolution of the AS when designing spatial filters. Finally, several methods of HRTF interpolation are reported in part 4.2.4. The statements concerning the application of those techniques are the object of section 4.4.

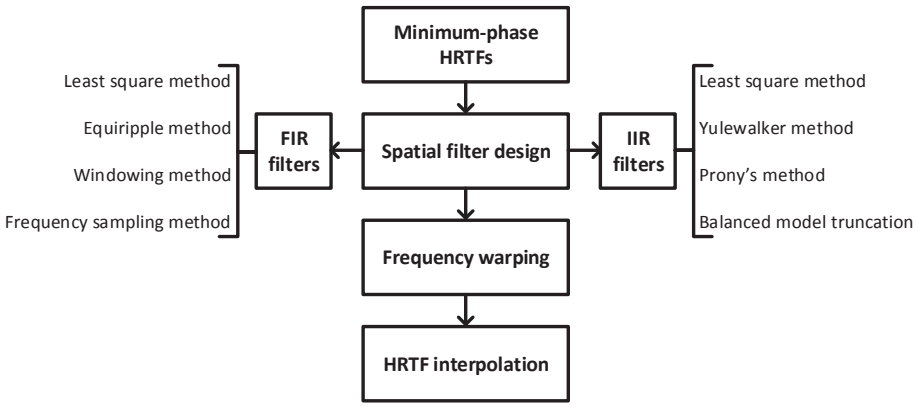


Figure 4.2 – The consecutive studied methods for the management of spatial filters.

4.2.1 Minimum-phase property

In this thesis, the database from the U.C. Davis CIPIC Interface Laboratory [4] is used. It provides the HRIRs of 45 subjects (including 2 KEMAR, one with small pinna, one with large pinna) at 25 different azimuths and 50 different elevations. The HRTFs from the KEMAR with large pinna, measured at 0° elevation and several azimuths in the frontal plane are taken on. These HRTFs are actually DTFs, since the response of the loudspeaker, microphone and room have been removed, and no ear canal resonance is present. The sampling frequency is 44.1 kHz. Thus, the HRIRs have been downsampled to the working sampling rate of 16 kHz. Commercial applications are permitted as long as a written acknowledgment is sent to the CIPIC.

HRTFs have been found to look close to *minimum-phase systems* [101], i.e. their excess phase (the phase shift in excess of the minimum-phase shift) is almost linear and can be approximated with a pure delay. It is therefore possible to separate the insertion of the ILD and spectral cues (using filters with minimum phase) and the introduction of the ITD (via a pure delay). Another advantage is that the minimum-phase version of the HRIR concentrates the maximum of the energy on the first samples, which is beneficial both for filter order reduction

and HRTF interpolation [102] (see part 4.2.2 and 4.2.3). Hence, this representation of the HRTFs is chosen. The 2 steps to get the minimum-phase representation of a HRTF pair are the following: extract the minimum-phase function of both HRTFs, and determine the ITD. For the extraction of the minimum-phase function, one usually resorts to the procedure by Oppenheim, reported in [35]:

The window w is defined such that:

$$w[m] = \begin{cases} 1 & \text{for} \quad m = 1 \text{ or } m = \frac{M}{2} + 1 \\ 2 & \text{for} \quad m = 2, ..., \frac{M}{2} \\ 0 & \text{for} \quad m = \frac{M}{2} + 2, ..., M, \end{cases} \tag{4.1}$$

where M states for the length of the HRIR (128 samples).

Then, the windowed cepstrum on both sides $C_{L,\theta}$ and $C_{R,\theta}$ are computed as follows:

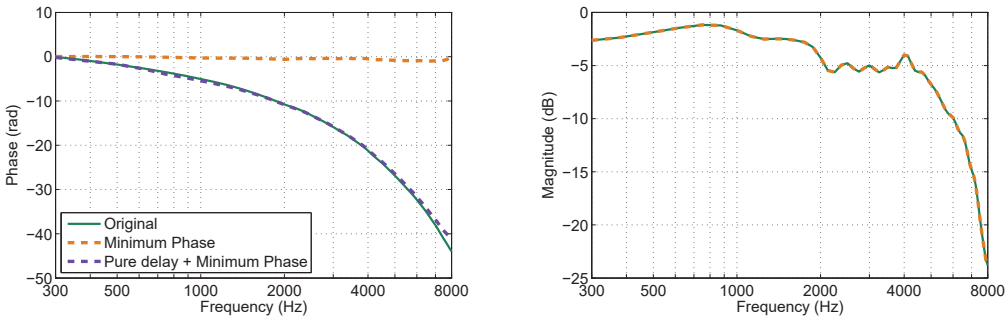
$$\begin{cases} C_{L,\theta}[m] = w[m] \times \mathfrak{F}^{-1}\left\{ \log\left|\mathfrak{F}\{h_{L,\theta}[m]\}\right| \right\} \\ C_{R,\theta}[m] = w[m] \times \mathfrak{F}^{-1}\left\{ \log\left|\mathfrak{F}\{h_{R,\theta}[m]\}\right| \right\}, \end{cases} \tag{4.2}$$

with $\mathfrak{F}$ and $\mathfrak{F}^{-1}$ denoting the Fourier and inverse Fourier transforms, and $h_{L,\theta}$ and $h_{R,\theta}$ are the left and right HRIRs. Finally, the minimum-phase versions of both HRIRs $h^{\min}_{L,\theta}$ and $h^{\min}_{R,\theta}$ are derived as follows:

$$\begin{cases} h^{\min}_{L,\theta}[m] = \mathfrak{F}^{-1}\left\{ \exp\left(\mathfrak{F}\{C_{L,\theta}[m]\}\right) \right\} \\ h^{\min}_{R,\theta}[m] = \mathfrak{F}^{-1}\left\{ \exp\left(\mathfrak{F}\{C_{R,\theta}[m]\}\right) \right\}. \end{cases} \tag{4.3}$$

To determine the ITD, the method reported in [209] has been first tested. It consists in extracting the pure delay component of each HRIR and computing the difference. To do so, one looks for the first sample of the *impulse response* (IR) that is higher than 5% of the maximal amplitude of the HRIR. Since this technique has led to underestimated ITDs, it has been finally chosen to directly derive the delay between both HRIRs, looking for the time lag that maximizes the cross-correlation function, as described by Kistler and Wightman [121].

Figure 4.3 shows an example of a minimum-phase function extraction. The HRTF of the contralateral ear at -80° has been taken on and its minimum-phase representation has been derived. Figure 4.3A displays the phase spectrum of the original and minimum-phase versions of the IR. It can be noticed that the minimum-phase transformation leads to an almost null phase. Also, it is evidenced that the additional pure delay allows to recover a great deal of the original phase. The magnitude is effectively left unchanged by the minimum-phase extraction (Figure 4.3B).



(a) (b)

Figure 4.3 – Comparison between the original (green) and minimum-phase (orange) versions of the contralateral ear HRTF at -80°. A: Phase spectrum, with the added pure delay + minimum phase (purple). B: Magnitude spectrum.

4.2.2 Filter design

Filtering with HRTFs implies to design adequate digital filters, which combine an accurate amplitude (for a correct reproduction of the ILD and spectral cues) and an accurate phase (for a correct rendering of the ITD). The specifications introduced in Chapter 1.4.3 require to develop the simplest filters as possible, i.e. the filters with a lowest order and a straightforward structure.

Finite impulse response filters

The 128-point HRIRs that come from the CIPIC database correspond to *finite impulse response* (FIR) filters of order 127. Obviously, smaller orders are required in order to significantly decrease the computational cost of such filters and the number of coefficients to be stored. For instance, Kulkarni and Colburn [128] report that a 64^{th} order FIR filter is sufficient to render most of the spatial information. Hartung and Raab [94] claim that the localization performance is not affected by the use of 48^{th} order FIR filters, and filters of order 32 only imply minor divergence. In this thesis, it is expected to reach even lower order of FIR filters for several reasons. First, the previously reviewed studies design spatial filters to cover the entire 3D space around the listener, while only the HRTFs in the FHP are considered here, for which the HF spectral content is of less importance. Second, previous studies perform localization test relying only on the auditory cues. In the thesis application, the visual cue is of great help to match the spatialized sound with the speaker's real location. Finally, the sampling rate is 16 kHz, which is 2.8 times smaller than the original 44.1 kHz sampling frequency. All the frequency components between 8 and 22.05 kHz are not reproduced, so the filter design does not need to consider them.

4 methods have been considered for the design of FIR filters:

- The *least square* (LS) method, where the goal is to minimize the approximated magni-

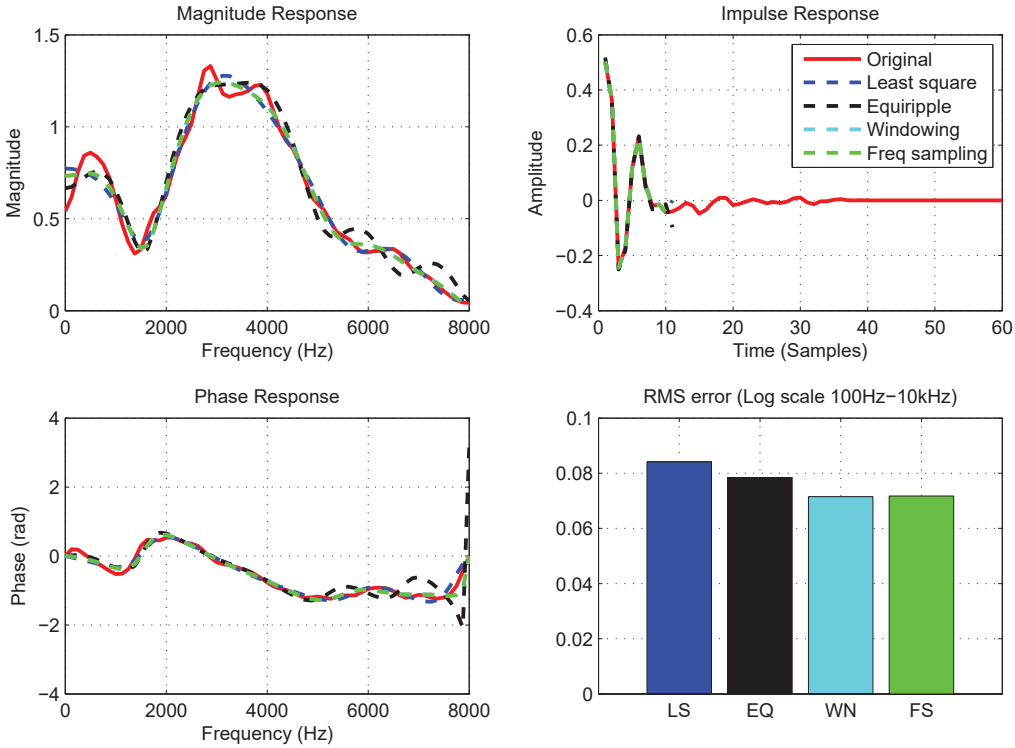


Figure 4.4 – Results of the implementation of FIR filters with the LS (blue), EQ (black), WN (light blue), and FS (green) methods, compared to the original HRIR (red). The upper left corner represents the magnitude response of the TF, the lower corner is the corresponding phase response. The upper right corner depicts the IR, and the lower right corner shows the RMS error between the original and approximated TFs, computed on a logarithmic scale between 100 Hz and 10 kHz. Taken from [41].

tude and phase responses of the designed filter relative to the original one in a least square sense [157],

- The *equiripple* (EQ) method that puts exactly the same number of ripples of minimized amplitude inside the pass-bands and stop-bands (only even order are possible for this technique) [154],

- The *windowing* (WN) method that consists in truncating the original IR. A rectangular window is used, since it was reported in [102] that it performs better than some other classical windows, such as the Hamming window, despite the inherent Gibbs phenomenon (ripples around amplitude response discontinuities) [155],

- The *frequency sampling* (FS) method, where the magnitude and phase responses are sampled at regular frequency intervals. Then, the corresponding IR is recovered in the time domain [156].

The results of the implementation of the 4 different methods is depicted on Figure 4.4, with the example of the HRTF of the contralateral ear at 45°, and an order equal to 10. The magnitude, phase and IRs are shown, as well as the *root mean square* (RMS) error between the original and approximated magnitude responses. It is computed on a logarithmic scale from 100 Hz to 10 kHz. It must be noted that this error is indicative and that there is no evidence that a correlation exists between this value and the perception of the different filters by subjects. The HF content is prominent for the ILD and spectral cues. On the contrary, the frequency-dependent phase response is of less importance, because the inserted pure delay corresponding to the ITD is dominant. The behavior of the minimum phase HRIR is highlighted, i.e. a great deal of the energy is concentrated on the 7 first samples. All methods smooth the original magnitude response and limit the IR to 11 samples (i.e. the 11 coefficients of the filter). Notice that the equiripple method leads to the strongest approximation error in the HFs, especially above 5 kHz, for both the magnitude and phase spectra. The overall RMS error indicates that the WN and FS methods yield the bast approximations. Therefore, they are preferred in the following (see part 4.4).

There are several advantages to implement FIR filters: they are always stable, they can be designed with a linear phase and the design methods are simple. A prominent advantage of FIR filters is the fact that they are easy to interpolate, as seen in part 4.2.3. FIR filters can also be implemented in the frequency domain, using the so-called *overlap-add method*, which consists in multiplying the magnitude and phase spectra of the signal directly with those of the filter.

Infinite impulse response filters

Infinite impulse response (IIR) filters can be computationally more efficient than FIR filters. On the other hand, they may lead to stability problems, especially when used in systems with a fixed-point resolution, where error accumulation and propagation can yield to divergence issues. A P^{th} order IIR filter has $2P+1$ coefficients while a Q^{th} order FIR filter has $Q+1$. In terms of memory storage of the coefficients, IIR filters are more interesting than FIR filters as long as:

$$P \leqslant \frac{Q}{2}. \tag{4.4}$$

When using the classical direct form II transposed structure [153], a Q^{th} order IIR filter leads to $2Q+1$ multiplications and Q additions, i.e. a total $3Q+1$ operations. A P^{th} order FIR filter requires $P+1$ multiplications and P additions, which gives a total of $2(P+1)$ operations. In terms of computational cost, IIR filters are more interesting than FIR filters as long as:

$$P \leqslant \frac{2}{3}Q. \tag{4.5}$$

For the development of the BSA, RAM saving (related to the computational cost) is more critical than ROM saving (relative to the coefficient storage). Therefore, only the second condition has been considered. In the literature, it has been reported that 48th order IIR spatial filters provide equivalent satisfying results than 72th FIR filters [211]. In this case, both FIR and IIR representations lead to the same computational cost, so there is no interest of using IIR filters. Another study by Asano *et. al* states that an order of 30 is enough to reproduce the elevation information, but that front/back confusions are more important compared to the use of the original HRTFs.

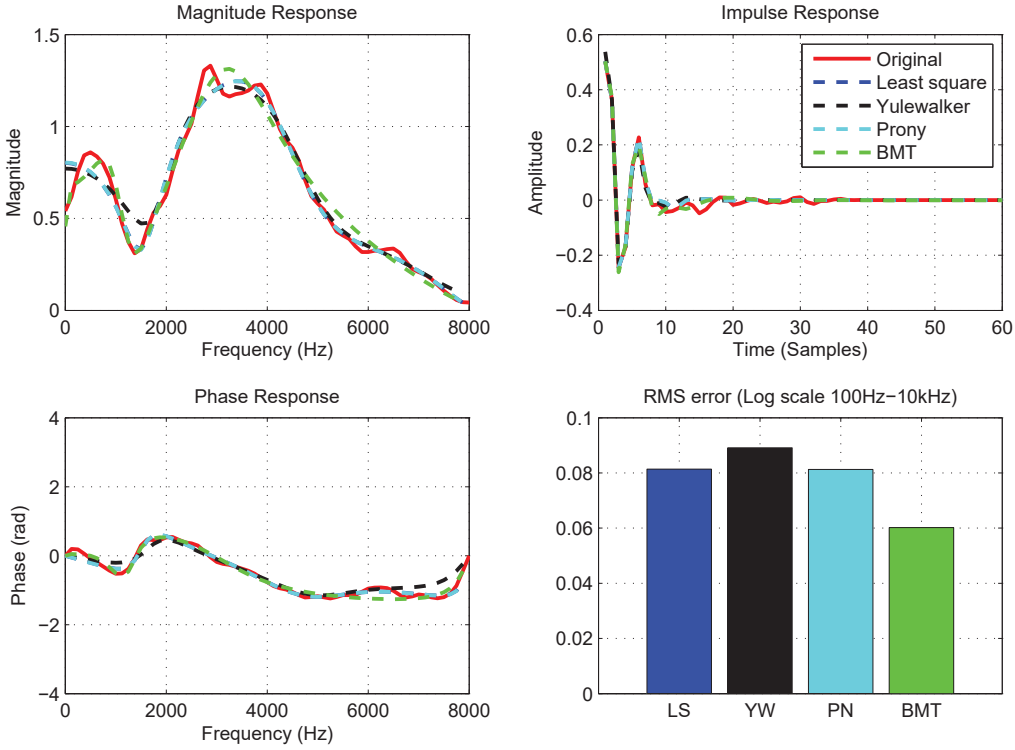


Figure 4.5 – Results of the implementation of IIR filters with the LS (blue), YW (black), PN (light blue), and BMT (green) methods, compared to the original HRIR (red). The upper left corner represents the magnitude response of the TF, the lower corner is the corresponding phase response. The upper right corner depicts the IR, and the lower right corner shows the RMS error between the original and approximated TF, computed on a logarithmic scale between 100 Hz and 10 kHz. Taken from [41].

As for the FIR filters, 4 methods have been considered for the design of IIR filters:

- The LS method that minimizes in a least square approach the approximated magnitude and phase of the filter relative to the original TF [158],
- The *Yulewalker* (YW) algorithm that performs a least square fit of the magnitude response only. This method is used in several studies such as [101, 102, 211], where it is

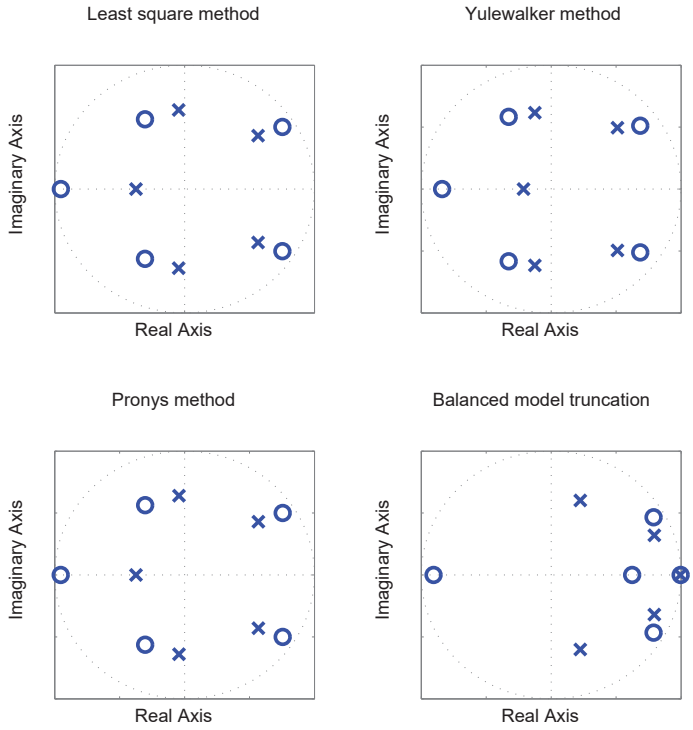


Figure 4.6 – The unit circle showing the repartition of the poles (crosses) and zeros (circles) of the filters designed with the 4 reported methods.

common to modify the algorithm so that more weight is given to the fit in the LFs [160],

- The *Prony's* (PN) method, based on a least square fitting and linear prediction, which is known to be very efficient for modelling short IRs [159],

- The *balanced model truncation* (BMT), which is based on a state-space representation of the original TF that is then truncated to achieve a filter order reduction. Huopaniemi and Karjalainen [102] manage to decrease their spatial filter order down to 10 thanks to this method. Mackenzie *et al.* [146] even state that "*no model derived with Prony's or the Yule-walker method is better than that computed with BMT*".

Figure 4.5 shows the implementation of these 4 methods for the HRTF of the ipsilateral ear at 45°, and an order equal to 5 (to get the same number of coefficients as the previous FIR design example). With IIR filters, a complete IR is obtained, although the approximation is a bit far from the original HRIR upper the 13^{th} sample. The RMS errors are on the same order of magnitude as the ones obtained with the FIR design (Figure 4.4), except for the BMT design that brings about a significant smaller error. Therefore, it might be possible to reach even lower orders thanks to this method, which would be in agreement with the results of [102].

However, the risk of instability is high with the BMT method, as depicted on Figure 4.6, which shows the repartition of the poles and zeros on the unit circle for the 4 designed filters. To ensure filter stability, all the poles have to be located inside the circle. On the contours, the system is still stable but it is also substantially sensitive to truncation error under a fixed-point resolution. With a pole located on the contour of the unit circle, the BMT design is a risky method to implement.

4.2.3 Frequency warping

In the design of digital filters, the fitting of the approximated TF to the original one is usually done on a linear frequency scale. This means that the same amount of error is minimized in the 0-1000 Hz band as in the 5000-6000 Hz band. When digital filters are used for audio applications, it is relevant to model the frequency resolution of the ear, i.e. to look for a fit of higher resolution in the LFs rather than in the HFs. It is possible to apply such a LF weighting during the design phase, e.g. in the Yulewalker algorithm, as already indicated in the previous part. Another technique consists in processing the TF prior to the filter design, so that no change is required in the design stage. This can be achieved by frequency warping.

Frequency warping operates a spread of the LFs toward the HFs. That means that the LF components of the TF are expanded, whereas the HF spectrum is compressed on a narrower interval. It can be viewed as the opposite of the frequency compression performed in HAs (Chapter 1.1.2). An extensive review on frequency warping and its audio applications can be found in [92]. Basically, the warped version of the IR is computed by filtering it with a concatenation of first-order all-pass filters. Those filters are monitored by a parameter b between -1 and 1 (bilinear conformal mapping). When b is positive, the LFs are expanded (filter design application), when it is negative, the HFs are spread (HA application). The higher the value of b, the more the warping. Figure 4.7 depicts this transformation with the HRTF of the ipsilateral ear at 45° when $b = 0.5$. One clearly notice the LF spreading, that moves the notch at 1.4 kHz on the original TF up to 3.6 kHz in the warped version. The same appears with the original peak at 4.3 kHz that goes to 7.5 kHz. With such a shape, the filter design mode focuses automatically more on the LFs and minimizes the approximation error in this part of the spectrum. Note that warping must be applied with care, because too much expansion of the LFs would deteriorate the approximation of the ILD and spectral cues that are located in the mid and HFs.

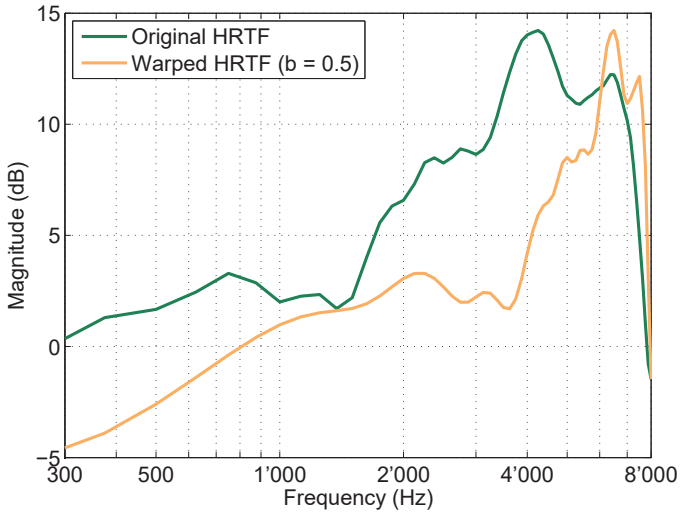


Figure 4.7 – Example of frequency warping, representing the original HRTF (green) and its warped version (orange), with $b = 0.5$. Taken from [41].

Figure 4.8 – Representation of the warping/unwarping process in the time domain. The green curve is the original HRIR, the orange curve is the warped version, and the dashed light green curve shows the HRIR recovered via the unwarping process. Taken from [41].

One important property of the warping process is that it transforms a finite IR (e.g. the 128-point HRIRs from the CIPIC database) to an infinite IR. This response has to be truncated to $P+1$ samples, where P is the order of the resulting filter in the case of a FIR filter design. Obviously, the resulting FIR or IIR filter coefficients cannot be directly applied because they correspond to a deformed TF. Therefore, *unwarping* must be performed so as to recover the original frequency response. Unwarping is exactly the same operation as warping, except that

the b parameter is taken as the opposite of the one used for warping. Figure 4.8 shows the original HRIR of the ipsilateral ear at 45°, its warped version with $b = 0.5$, and the unwarped version ($b = -0.5$) that allows to recover the original HRIR.

Unwarping can easily be applied for FIR filters, since the IR is made of the filter coefficients. Hence, once the warped coefficients are computed, unwarping on these coefficients is performed, and the filtering is processed using the direct form II transposed structure. However, this is quite a bit more complex for IIR filters. Such filters can be designed using a warped TF but it makes no sense to apply unwarping on the numerator and denominator coefficients. Instead, a specific filtering procedure that keeps the warped coefficients has been developed by Karjalainen and Härmä in C code [91] and translated in Matlab code for the needs of this thesis. This procedure is more expensive in terms of computational cost, e.g. it took 8.5 times more under Matlab, for an 8^{th} order IIR filter based on a 128-sample HRIR.

In the literature, *warped FIR* (WFIR) are hardly ever used because they do not allow a significant order reduction. On the other hand, *Warped IIR* (WIIR) filters have been found to convey valuable results, despite the increased computational time. A discrimination subjective test reported in [101] establishes that 75% of the 8 tested subjects are not able to differentiate the original and approximated HRTFs, with orders of 40, 25 and 20 for the FIR, the IIR and WIIR filters respectively. In [102], Huopanoemi *et. al* conclude that a WIIR filter of order 16 is sufficient to render most of the perceptual features of HRTFs.

4.2.4 Interpolation

HRTF/HRIR interpolation is a common topic in 3D audio. This operation is required when one wants to spatialize a source in any direction, while only a limited set of HRTF pairs are available. The underlying goal is also to minimize the HRTF database. A typical use case is when it is desired to render an artificial motion of a sound source, e.g. a speaker moving in a classroom. Performing HRTF interpolation allows to spatialize a sound source in an intermediate location, for which no HRTF pair is available. This is not a trivial task and various methods have been reported to this end. Table 4.1 reports some methods in the time domain, namely the linear interpolation of the HRIRs, the pole-zero linear interpolation, and the interpolation based on the common acoustical poles and zeros. Finally 2 methods that can be used in both the time and frequency domains are introduced in Table 4.3: the inverse distance weighting and the spherical splines interpolation. The “dimension” column stands for the spatialization dimension (2D or 3D) that is reachable with each procedure.

Method	Dimension	Principle	Remark(s)
HRIR linear interpolation [3, 168]	2D	1. Separation of the 2 closest HRIRs into their minimum phase version and a pure time delay, 2. Linear interpolation of both components.	Easiest method.
Pole-Zero linear interpolation [195]	2D to 3D	Perform linear interpolation for each pole and zero with the neighbouring poles and zeros of known HRTFs. This is not trivial and there is no guarantee that the interpolated filter is stable. Some filter structures (e.g. Kalman) can facilitate the processing.	Require an IIR-filter implementation.
Common acoustical poles and zeros-based interpolation [195]	2D to 3D	1. Time alignment of the original IR, 2. Determination of the common acoustical poles and zeros to obtain an IIR filter with direction-independent poles, direction-dependent zeros, and residues, 3. Linear interpolation of the zeros and residues. The poles can be kept constant.	Require a IIR-filter implementation.

Table 4.1 – HRTF-interpolation methods working in the time domain. Taken from [41].

Method	Dimension	Principle	Remark(s)
Interpositional Transfer Function-based (ITPF)-based interpolation [74]	3D	The HRTFs of each location on a sphere are approximated by the linear combination of the TFs of the 3 vertices of a triangle containing the desired point. The original HRTF of the nearest point is used while one resort to an approximated version of the HRTF (called IPTF) for the 2 others.	2 main advantages: 1. Only 3 HRTFs required, 2. The 2 IPTFs can be made low orders.
emphPrincipal component analysis (PCA)-based interpolation [31, 118]	3D	1. Compression of the frequency domain magnitude components of HRTFs using PCA, 2. Interpolation applied on component weights using a certain method (spherical splines in [31], or based on some rational functions [118]), 3. Decompression of the interpolated PCA weights to recover the desired HRTFs.	

Table 4.2 – HRTF-interpolation methods working in the frequency domain. Taken from [41].

4.3 HRTF dynamic limitation

This second part reports the scientific contribution of this thesis concerning the implementation of a BSA for aided HI subjects. The principle of the HRTF dynamic limitation is detailed, as well as the results obtained from a psychoacoustic study.

Method	Dimension	Principle	Remark(s)
Inverse distance weighting interpolation[32]	3D	1. Separation of the DTFs into log-mag and log-phase components, 2. Derivation of the great-circle distance for the 4 positions closest to the desired one, 3. Computation of the weights as the inverse distances, 4. Derivation of the interpolated magnitude and phase as the weighted sum of the 4 closest HRTF locations. ITD can be derived with interpolation of pure-time delaies or mathematical formula (e.g. Woodworth's formula).	Work in the time domain with FIR filters only (bilinear interpolation).
Spherical spline interpolation [94, 240]	3D	The overall principle is to interpolate a given HRIR or HRTF using the complete available dataset. The method is based on the interpolation with cubic splines and the estimation of the second derivatives of the HRIR/HRTF dataset, under certain hypotheses.	More complicated and time consuming than linear interpolation, but give better perceptual results.

Table 4.3 – HRTF-interpolation methods working in both time and frequency domains. Taken from [41].

4.3.1 Spatialization for hearing-impaired subjects

The literature dealing with spatialization applied to HI listeners is scarce and recent. The concept is suggested in a patent of Oticon [90] that proposes to implement speech spatialization with WMS. Nevertheless, no concrete method is described. Ohl *et al.* [184], from Oticon, have shown that HI subjects are sensitive to HF spectral cues allowing externalization, despite variable performance among listeners and a reduced sensitivity compared to NH subjects. Externalization in HI listeners is the major topic of the article of Boyd *et al.* [24]. It is reported that the 14 tested HI subjects suffer from a contracted perception of externalization, due to their decreased sensitivity to pinna cues. The authors hypothesize that HI listeners put a greater emphasis on the ITD, ILD and DRR rather than on spectral cues to localize sound sources, which also explains the loss of externalization. Whitmer *et al.* [244] have studied the perception of virtual auditory images width in a group of 35 NH and HI participants. Variations of the IC has been used to control the width. The outcomes show that there is no correlation between the perceived width of the sound images and HRL. However, age has a significant effect, i.e. older subjects report a narrower variation of width with changes of the IC, and some of them even perceive no difference. This could provide with an insight on a possible lower perception of spatialization in old HI subjects.

Despite the use of customized HRTFs, it seems that the rendering of a realistic externalization is difficult in HI people. Madjak *et al.* [147] attempt to apply spectral warping. The frequency content between 2.8 to 16 kHz is linearly warped to a new range between 2.8 and 8.5 kHz. The objective is to establish whether an alternative encoding of the spatial cues is possible on NH subjects, thanks to a long-term training. The outcome on 6 participants shows the

inefficiency of this method, yielding to unnatural ITDs and ILDs that the listeners are not able to get used to. Interestingly, the subjects reach better performance after a period of training with band-limited HRTFs (cut-off frequency at 8.5 kHz). This is in agreement with the fact that HI people are capable of localizing sound sources despite their limited access to the content in HFs (see Appendix B.1.2). Mueller *et al.* [173] (in collaboration with Phonak) evidence that it is possible to provide an artificial spatial perception in HAs for 12 NH subjects, but they do not study the internalization and the perception of width. The next section goes further in the idea to apply spatialization on a pair of HAs, while taking into account the properties of the impaired AS.

4.3.2 Spatialization on hearing aids

Principle

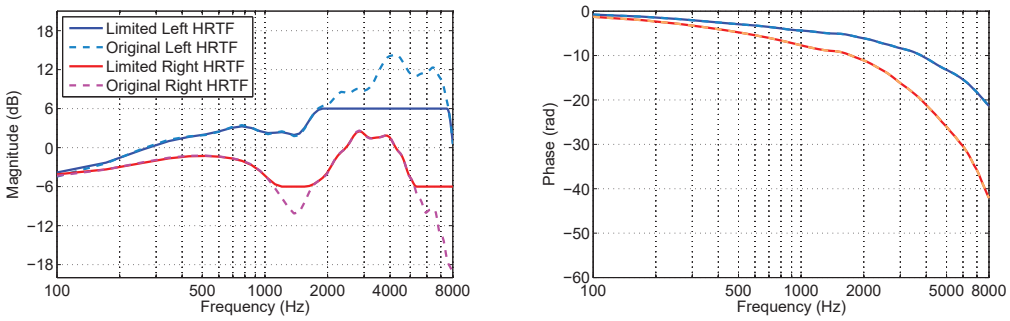
The implementation of HRTF-based filters on a pair of HAs may require prior modifications of the HRTF spectra. It is known that HI subjects suffer from a reduced auditory dynamic range (the recruitment phenomenon, see Appendix B.1.2). A limitation of the magnitude response of the HRTFs could be applied, in order to prevent the spatial filters from pushing the audible speech outside the range of audibility. However, it is expected that excessive limitation would be perceivable and would lead to a distorted spatial effect, as it yields a reduction of the ILD and some distortions of the spectral cues.

The suggested procedure of dynamic limitation is performed offline on the minimum-phase version of the HRIRs. Only the magnitude components of the HRTFs are affected by this processing, whereas the phase is preserved. The principle can be viewed as a limiter acting in the frequency domain, cutting and flattening each magnitude component that goes beyond a certain range. Limitation, rather than compression, has been chosen because previous internal and informal tests has shown that limitation provides less perceived distortion than compression. This can be explained by the fact that compression affects a greater part of the HRTF magnitude than limitation, resulting in more perceivable effects. On the contrary, limitation only concerns some local parts of the spectrum.

Given a desired dynamic range Δr in dB, the maximum and minimum gains at both ears are $\pm\frac{\Delta r}{2}$. The limitation is done symmetrically relative to 0 dB to ensure the lowest possible amplification or attenuation in both the ipsilateral and contralateral ears. This can be formulated as follows:

$$|\overline{H_\theta(f)}|_{\text{dB}} = \begin{cases} +\dfrac{\Delta r}{2} \text{ if } |H_\theta(f)|_{\text{dB}} > \dfrac{\Delta r}{2} \\ -\dfrac{\Delta r}{2} \text{ if } |H_\theta(f)|_{\text{dB}} < -\dfrac{\Delta r}{2}, \\ |H_\theta(f)|_{\text{dB}} \text{ otherwise} \end{cases} \tag{4.6}$$

where θ is the azimuth of the HRTF, $|\overline{H_\theta(f)}|_{\text{dB}}$ represents the limited version of the magnitude component $|H_\theta(f)|_{\text{dB}}$ given in dB.



(a) (b)

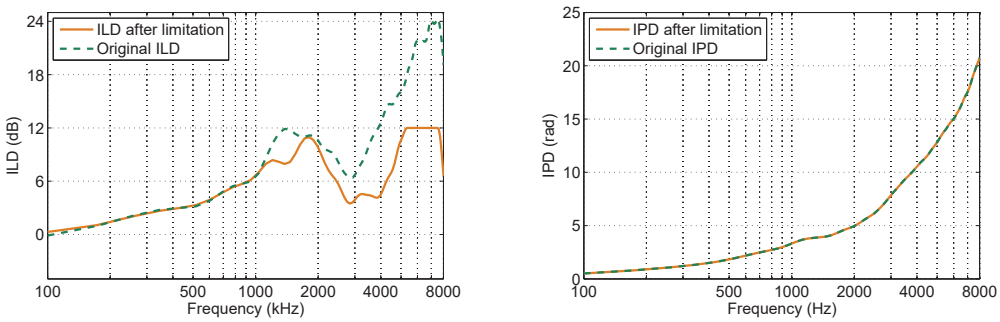
Figure 4.9 – Effect of a 12 dB dynamic range limitation on the magnitude of a pair of HRTFs at 45°. The dashed lines are for the original HRTFs and the solid lines represent the limited HRTFs. The HRTFs of the left ear are in dark/light blue. The HRTFs of the right ear are in red/orange. Taken from [51].

To ensure that the phase is not altered by the processing, the limitation is applied only on the magnitude of the HRTFs. The phase is saved before the processing, and combined with the new magnitude obtained after limitation. Therefore no perceivable effect could be due to ITD distortions. The distortion of the ILD and spectral cues is expected to cause some audible artifacts, such as source centering and internalization as the amount of limitation augments. Figure 4.9 shows an example of dynamic range limiting applied on a pair of HRTFs for a sound source located at 45°, with a 12 dB dynamic range. Figure 4.9A depicts the corresponding modification of the magnitude. The preservation of the phase is evidenced on Figure 4.9B. There is no limitation occurring below 1 kHz since the gains are inside the allowed dynamic range. On the contrary, in the HFs, both the ipsilateral and contralateral TFs are alternatively or simultaneously limited. This impacts the HF monaural cues. The resulting binaural cues are presented on Figure 4.10, i.e. the ILD (Figure 4.10A) and IPD (Figure 4.10B). No change occurs for the IPD. However, the ILD is affected by the limitation in the HFs (i.e. when it is used by the AS) and it decreases by 12 dB in certain frequency area. The perceptual effect of this distortion might be a centering of the sound image in the FHP. In order to confirm this hypothesis and determine the maximum amount of limitation that can be applied before subjects perceive audible artifacts, an extensive psychoacoustic evaluation has been conducted, and is presented hereafter.

Psychoacoustic evaluation

Subjects, stimulus and setup

39 inexperienced NH listeners, aged between 18 and 26 years (median age = 21 y.o.), took part in the experiment. Their bilateral hearing thresholds at 0.5, 1.5 and 4 kHz were checked with



(a) (b)

Figure 4.10 – ILD (A) and IPD (B) resulting from the dynamic limitation depicted on Figure 4.9. The original cues are in green dashed line and the modified ones are in orange solid lines.

an audiometer before starting the experiment, so as to ensure normal hearing (i.e. hearing thresholds lower than 20 dB HL). One participant did not fulfill this requirement. Hence, the corresponding results were discarded and the sample size was finally 38. All the listeners were paid for their participation.

The stimulus used was a 2-second sample of the English male speech from the EBU SQAM CD [242]. Previous informal tests aiming at comparing the use of a male and a female voice concluded that the minimal dynamic range before getting subjective audible effects was smaller in the case of the male speech, which was therefore chosen as a worst case. The sampling frequency was set to 22.05 kHz, which is a typical rate in Phonak HAs. The stimulus was filtered with different pairs of HRIRs corresponding to various azimuths in the FHP. These HRIRs were downsampled from 44.1 kHz to 22.05 kHz and converted into 128-tap FIR filters.

The picture of Figure 4.11 shows the global setup. The stimuli were presented through headphones (Beyerdynamic DT770) at 65 dB(A) (RMS value measured for a stimulus filtered with the 0° pair of HRTFs and no limitation applied) in the EPFL anechoic chamber. The frequency response of the headphones had been previously estimated by repeated measurements on a *head and torso simulator* (HATS B&K type 4128), and compensated in the playback stimuli. The participants were facing the touch screen of a laptop running the *graphical user interface* (GUI) developed for the experiment. This laptop was connected to a digital audio interface (M-Audio Fast Track Ultra 8R) that played the sound in the headphones.

Test design and procedure

A *constant reference duo-trio discrimination test*, as described by Lawless and Heimann [131, Chap. 4], has been conducted in this study. The listeners were played 3 samples, among which the first was the reference and one of the 2 consequent samples was identical to the reference, while the other was different. The subject had to state which sample was similar to the reference. The reference was a stimulus filtered with a certain pair of non-limited HRTFs, whereas the different sample was filtered with the pair of limited HRTFs for the same

Figure 4.11 – Picture of the psychoacoustic test setup.

azimuth and a given limited dynamic range. A run consisted in listening to the 3 samples twice and choosing which of the 2 last was identical to the reference. The double listening was demanded in order to decrease the risk of a random answer due to inattention, while ensuring that all subjects had the same number of listenings. The participants were instructed to focus on the perceived incidence direction of the spatialized source, on the externalization and on the frequency balance.

The succession of runs was governed according to an adaptive procedure called the *simple staircase* [131, Chap. 6]. A total of 9 azimuths θ (±80°, ±65°, ±40°, ±20° and 0°) and 13 dynamic ranges Δr (between 10 to 34 dB in 2 dB steps) were available. A flowchart of the whole procedure is shown on Figure 4.12. For each azimuth, the test began with a fixed dynamic range of 16 dB, then each correct answer (correct determination of the sample similar to the reference) led to a dynamic range increased by 2 dB (less limitation). Every false answer led to a dynamic range decreased by 2 dB (more limitation). The different azimuths were randomly selected during the test. For each azimuth, the stop condition was reached when the listener had given 2 right answers for a certain dynamic range. Then, the threshold for the corresponding azimuth was assumed to be the previous dynamic range for which the answer was correct. The test finished when the thresholds for the 9 azimuths were determined.

Results

Type-I and -II errors

Discrimination tests are subject to 2 main kinds of errors: the so-called type-I (false rejection) denoted α, and type-II (miss) errors denoted β. In the present study, the null hypothesis states

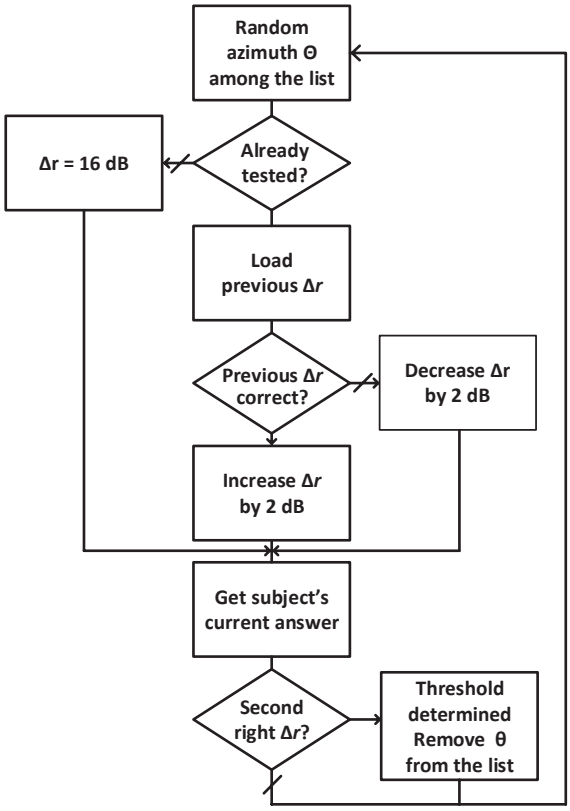


Figure 4.12 – The simple staircase procedure governing the psychoacoustic test.

that no difference is perceivable between the original and limited HRTFs. The type-I error corresponds to the rejection of the null hypothesis when it is actually true, which leads to the determination of a too high required dynamic range while more compression could be applied. On the other hand, the type-II error relates to the case where the null hypothesis is accepted whereas it should not be, i.e. the minimum dynamic range is underestimated and less limitation should be applied.

Lawless and Heimann [131, Chap.5] reported a formula that provides an estimate of the necessary sample size N, depending on the α and β value:

$$N = \left(\frac{Z_\alpha \sqrt{p_0 q_0} + Z_\beta \sqrt{p_A q_A}}{p_0 - p_A} \right)^2 , \tag{4.7}$$

where Z_α and Z_β are the Z-scores associated with the chosen levels of α and β, p_0 is the chance probability of the test ($q_0 = 1 - p_0$) and p_A is the proportion of desired correct answers (after chance correction) in order to establish the perceptual threshold ($q_A = 1 - p_A$). A value of 10% was chosen for the type-I and type-II errors, the associated Z-scores are equal to 1.828 according to the table of the standard normal distribution, and the chance probability of a

constant reference duo-trio test is 50%. Chance correction is performed by computing p_A using the Abbott's formula [131, Chap. 5]:

$$p_A = P_{\text{cor}} + p_0(1 - P_{\text{cor}}), \tag{4.8}$$

where P_{cor} is the proportion of correct answers (before chance correction) that defines the threshold, which was set to 50%, that is, the perceptual thresholds are the minimum dynamic ranges required so that half of the participants do not distinguish between the original and limited HRTFs for a given azimuth. It yields $P_{\text{cor}} = 0.5$, $p_A = 0.75$ and $N = 22.89$, which means that at least 23 subjects are necessary to determine the perceptual thresholds with an amount of type-I and type-II errors lower than 10%. It should not be inferred that a sample size of 38 listeners is surely enough, because of the adaptive processing. Indeed, not all the participants have tested the whole combinations of dynamic ranges and azimuths. Figure 4.13 depicts the number of tested listeners in the different conditions. This figure will serve to validate the hypothesis that the determined thresholds can be established with type-I and type-II errors lower than 10% in the next section.

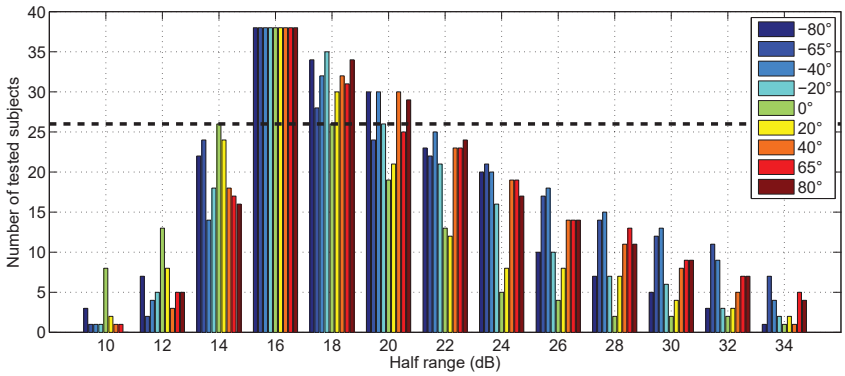


Figure 4.13 – The total number of tested subjects in the different combinations of dynamic ranges and azimuths. The black dashed line represents the minimum sample size of 23 participants that is required to get thresholds with less than 10% type-I and type-II errors. Taken from [41].

Threshold determination

Figure 4.14 displays the cumulative distribution functions of the individual thresholds for all azimuths, as a function of the dynamic range. The dashed line represents the 50% proportion chosen to define the perceptual thresholds in this experiment. Once this line is reached, one can state that at least half of the population cannot perceive any difference between the original and limited HRTFs.

A one-way within-subjects ANOVA has revealed a significant dependance between the thresholds and azimuths ($F_{8,333} = 1.966$, $p < 0.1$), meaning that it makes sense to establish one

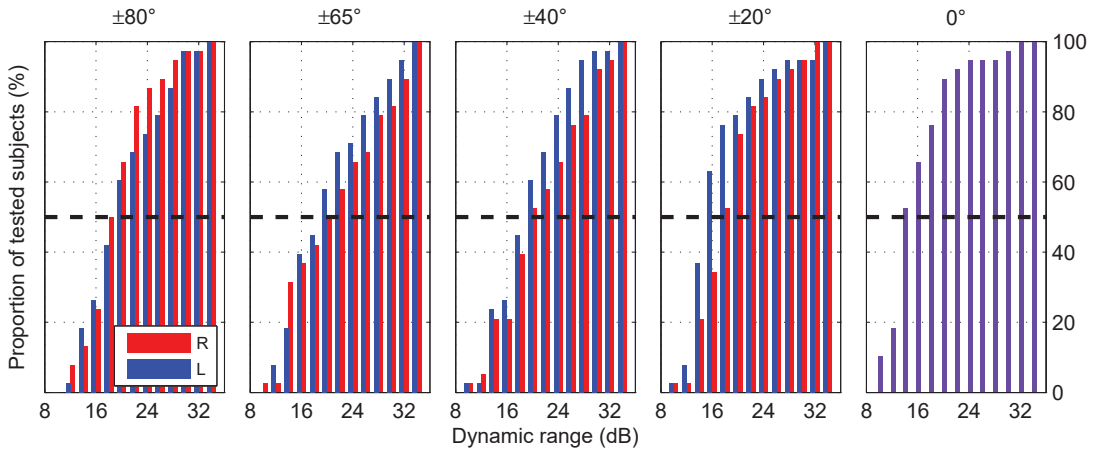


Figure 4.14 – The cumulative distribution functions of the individual thresholds as a function of the dynamic range. Results from the left azimuths are depicted in blue while those from the right azimuths are in red. The extreme-right panel shows the outcomes of the 0° azimuth. The black dashed line represents the 50% proportion that defines the perceptual threshold in this experiment. Taken from [41]

specific minimum dynamic range for each azimuth. Another one-way within-subjects ANOVA has shown no significant effect of the side on the determined thresholds ($F_{1,70} = 3.978$, $p = 0.4208$). Note that the data from the 0° azimuth have been excluded for this second analysis.

Table 4.4 summarizes the thresholds of minimum dynamic ranges determined from Figure 4.14 for all azimuths. The corresponding number of participants is indicated as well, and shows to be greater than 23 in all cases. It confirms that the results can be expressed with type-I and type-II errors lower than 10%.

Azimuth (°)	-80	-65	-40	-20	0	+20	+40	+65	+80
Minimum dynamic range (dB)	18	20	20	18	14	16	20	20	20
Number of tested subjects	34	24	30	35	26	38	30	25	29

Table 4.4 – Minimum dynamic ranges determined from Figure 4.14 for the different tested azimuths. The corresponding number of tested participants is given in the third line, according to Figure 4.13. Taken from [41].

Discussion

The reported psychoacoustic test introduces and investigates the concept of dynamic range limitation applied to HRTFs. The reported experiment enables to establish the minimum dynamic range that is necessary so that at least half of the 38 participants cannot hear any difference between the original and limited version of a HRTF pair. The total number of subjects is enough to claim type-I and type-II errors lower than 10 %. The minimum dynamic ranges are shown to depend on the azimuth, and increase when the spatialized sound source

moves from frontal to lateral azimuths. This can be easily associated with the fact that the ILD cue increases with the azimuths as well, following the same trend as the determined thresholds. Thus, more limitation can be applied in the frontal azimuths to yield a similar degree of ILD distortion as in the lateral azimuths.

The reported thresholds are valid for sources spatialized in the FHP, for a sampling frequency of 22.05 kHz, for NH subjects, and for generic HRTFs. It is likely that less dynamic limitation could be applied if a higher sample rate was used, if HRTFs from other spatial locations were considered (e.g. for other elevations, or to the rear of the listener), as well as if customized HRTFs were used. The greater emphasis that is given by the AS on the HF spectral cues might have resulted in a higher sensitivity to the distortions of the HFs caused by the limiting process. The suggested limited dynamics is tested on NH subjects, in order to validate the processing. It is expected that even smaller dynamic ranges would be possible for HI subjects, which can be considered as "bad localizers". Their limited access to HFs and the narrow bandwidth of HAs would certainly allow a more pronounced limiting. In addition, the presented minimum dynamic ranges are determined so that at least half of the NH listeners does not perceive any difference. There is a great chance that this proportion would be higher for HI subjects if the same minimum dynamic ranges were used.

4.4 Characteristics of the binaural spatialization algorithm

This part details the choice made for the BSA, after the processing and methods reported in the 2 previous parts. It describes the final spatial filters, and addresses the question about the HRTF interpolation, and the way how the BSA is implemented.

4.4.1 Final filters

This section discusses the final choices for the implementation of the BSA. After having reviewed the topics concerning the representation, filter design, limitation and interpolation of spatial filters in part 4.2, it is now possible to decide which strategies must be adopted and how the processing has to work with the BLA.

All the generic HRIRs are initially taken from the CIPIC database, selecting the KEMAR with large pinna at an elevation of 0°. The dynamic limitation is applied according to the results reported in the previous part. Note that the same thresholds have been kept even though the experiment was performed with a sampling rate of 22.05 kHz. Indeed, there is a great chance that the same dynamic ranges suit a 16-kHz sampling frequency. In a 5-sector resolution, the selected HRTFs correspond to azimuths at 0° (sector C), ±30° (sectors L1 and R1) and ±65° (sectors L2 and R2). In fact, these directions approximately match the central azimuth of each sector. When the BLA demands a 3-sector resolution, the lateral azimuths taken on for the spatial filters are at ±45°. Since the HRIRs at ±30° and ±45° have not been tested in the psychoacoustic test, it has been decided to keep a dynamic range of 20 dB, as it is the minimum

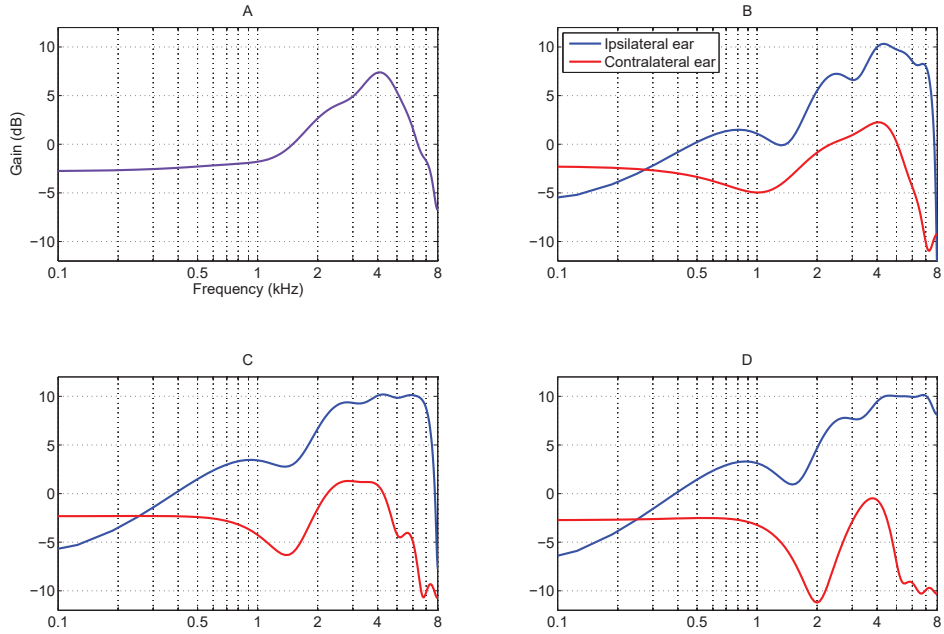


Figure 4.15 – Magnitude of the final filters implemented in the BSA. The ipsilateral ear is in blue and the contralateral ear is in red. All the selected azimuths are represented: 0° (A), ±30° (B), ±45° (C) and ±65° (D). Taken from [41]

dynamic range needed at 40° and the highest dynamic required for all the tested azimuths. In order to limit the number of spatial filters, the same pair of HRIRs is used, depending on whether the source is at θ or $-\theta$, i.e. only the left and right channels are inverted.

Minimum-phase FIR filters have been preferred for several reasons:

- Easier to interpolate,
- No risk of truncation error propagation,
- Simpler implementation structure,
- Guaranteed stability.

The resorts to FIR filters in the BSA makes the warping technique be irrelevant due to the rationales expressed in part 4.2.3. Moreover, the low sampling frequency does not really justify the use of warping, as the LF components are supposed to be sufficiently taken into consideration. Informal tests on a few experimented subjects have helped determine the minimum filter order needed, and the best design method among the 4 tested ones, so as to keep a good perception of spatial hearing with a sampling frequency of 16 kHz. When 2 or more methods have led to the same minimum order, the one that generates the lowest ripples around the compressed areas have been chosen. Since the interpolation process demands all

the filters to have the same order, the highest determined order among the different tested azimuths has been kept for all the others. Finally, the windowing method and a 10^{th} degree (i.e. 11 coefficients) are used for each spatial filter. The frequency response of the final filters implemented in the BSA are depicted on Figure 4.15. The magnitude is smoothed, due to the low order used to design the filters. The limitation with a dynamic range of 20 dB is clearly visible.

Table 4.5 summarizes the different full-band ITDs that are applied in complement to the minimum phase spatial filters and the corresponding number of sample shifts. These pure delay ITD values have been derived from the cross-correlation technique described in part 4.2.1. The implementation of the ITD is the object of the next part.

Azimuth (°)	0	30	45	65
ITD (μs)	0	227	363	542
ITD (sample shift at 16 kHz)	0	4	6	9

Table 4.5 – The ITD (in μs and in sample shifts) for the different azimuths used in the BSA. Taken from [41].

4.4.2 Interpolation

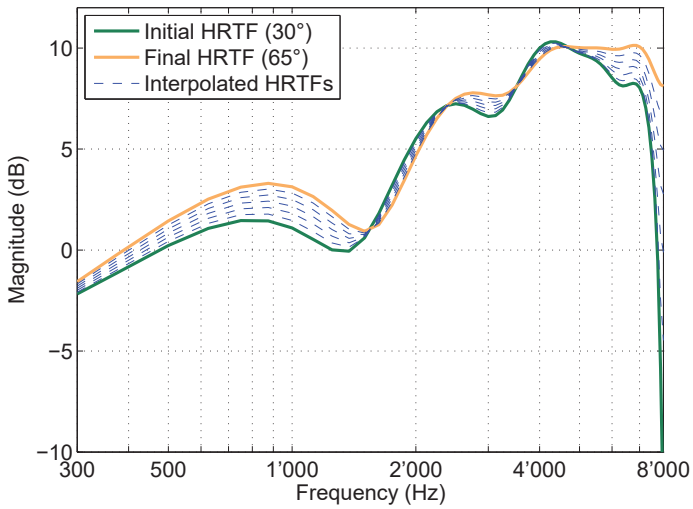


Figure 4.16 – Example of linearly interpolated HRTFs (dashed blue lines) between the initial HRTF of the ipsilateral ear at 30° (solid green line) and the final HRTF at 65° (solid orange line). Taken from [41]

The reported interpolation methods that use IIR structures in the time domain are not considered anymore. The use of temporal FIR filters also excludes the frequency-domain methods. Finally, only the linear interpolation and the spherical splines interpolation remain. This last method requires a large database of HRIRs in order to achieve good performance, while it is

requested to store as few filters as possible. Furthermore, it is significantly more complex than a linear interpolation. Therefore, the linear interpolation is chosen as a start, before deciding whether the rendering is sufficiently realistic, or if the spherical spline method should be considered instead.

Let $\widehat{h_\theta^{\min}}$ be the minimum-phase interpolated HRIR at the current frame k, and $h_{\theta_{\text{init}}}^{\min}$ and $h_{\theta_{\text{fin}}}^{\min}$ the available minimum-phase HRIRs at the initial and final direction θ_{init} and θ_{fin}. The current HRIR is computed as follows:

$$\widehat{h_\theta^{\min}}[k] = r\, h_{\theta_{\text{init}}}^{\min}[k] + (1-r)\, h_{\theta_{\text{fin}}}^{\min}[k], \quad \text{for } k = 1, \ldots, Q+1, \tag{4.9}$$

where r is the interpolation coefficient (expressed in number of frames), of which the value is detailed in the next part.

The same relation holds for the interpolated ITD $\widehat{\delta_\theta}$:

$$\widehat{\delta_\theta}[k] = r\,\delta_{\theta_{\text{init}}}[k] + (1-r)\,\delta_{\theta_{\text{fin}}}[k]. \tag{4.10}$$

Figure 4.16 shows an example of some different HRTFs interpolated between 30° and 65° (azimuths available in the BSA) with the previously described Equation 4.9 and 4.10.

4.4.3 Implementation

Despite the 8-ms analyse frame of the hardware, it is prominent to guarantee a continuous audio stream at the output of the system, that is, to produce a waveform without any discontinuity. When using the overlap-add technique, it is possible to process all the frames in an independent way, after having multiplied them with a certain temporal (analysis) window. When the filtering is done, the resulting frames have to be multiplied with the adequate synthesis window and temporally added together, so that the analysis and synthesis windows cancel out [69, Chap. 9]. This method is described on Figure 4.17A. One of its drawback is that it is redundant and thus computationally costly. Additionally, it is known to be advantageous when the signal to be filtered is long and when the filter is of high order. Therefore, it has been decided to resort to another technique that does not require any overlap and multiplication by windows. It is depicted on Figure 4.17B. Each successive frame is filtered using the initial conditions coming from the previous one, to avoid the initial convergence time delay and guarantee the continuity of the waveform. It allows to reduce the computational effort but requires the storage of 10 initial conditions (for a filter of 10^{th} order) at each frame,.

A tricky case occurs when a HRTF interpolation phase is ongoing. In this scenario, the filter coefficients change at each new incoming frame. In order to keep on using the aforementioned method, it must be assumed that the successive filters are sufficiently resembling, so that

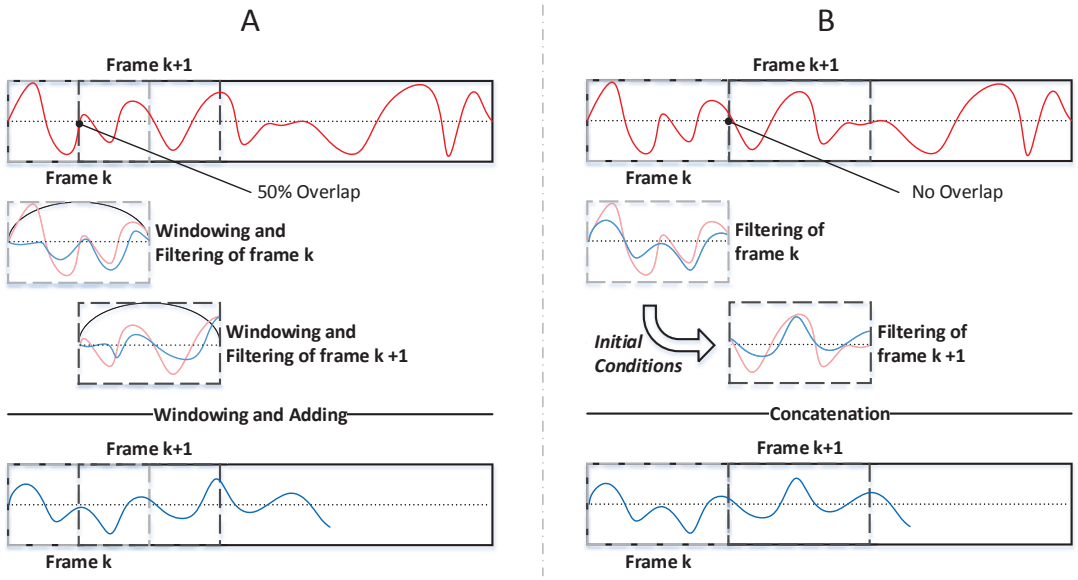


Figure 4.17 – Principle and comparison of the frequency-domain filtering (A) and the temporal-domain filtering (B) chosen in this thesis. The original signals are in red and the filtered signals are in blue. Taken from [41]

the initial conditions from the previous frame are still valid for the current one. This means that the interpolation cannot be done too fast, otherwise audible artifacts might occur. The standard interpolation time has been set to 1.2 second (i.e. $r = 150$ frames of 128 samples). It is sufficiently large for ensuring an adequate management of the initial conditions, and prevents listeners from hearing the existence of 5 sectors when the speaker is moving. When it has to be done faster (as explained below), it falls to 300 ms ($r = 38$). The sectorization is then quite a bit more audible.

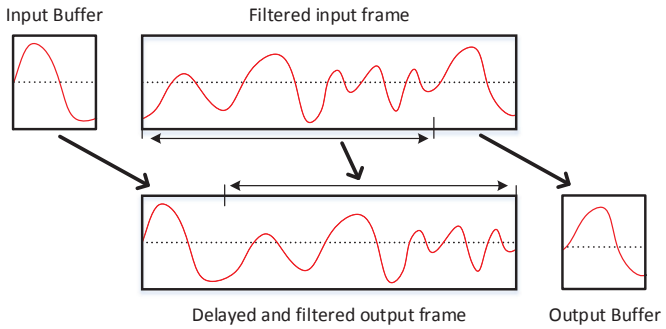


Figure 4.18 – Processing to introduce the adequate amount of ITD in the spatialized signal. Taken from [41]

As explained in part 4.2.1, the minimum-phase spatial filters only encapsulate the ILD and

spectral cues. The ITD cue has to be added separately, by applying the adequate pure delay reported in Table 4.5. If the speaker's voice has to be spatialized on the left (resp. right), the right (resp. left) channel is delayed by the suitable number of sample. The way how it is implemented is described on Figure 4.18. Once the filtering is done, the frame of the adequate channel is modified so that the few first samples are taken from the input buffer. The latter contains the adequate samples from the previous frame. Then, the few last samples are stored in the output buffer for the next frame.

3 types of transitions are implemented in the prototype, depending on what were the previous sector and resolution given by the localization block, and what are the current ones:

- The standard interpolation, conducted on 150 successive frames (1.2 s) when a 1-sector step occurred (e.g. going from the sector C to the sector L1), or when the resolution changes in the sectors L2 and R2,

- The fast interpolation, performed on 38 successive frames (300 ms) when a 2-sector step occurred (e.g. going from the sector C to the sector L2). The standard interpolation is also accelerated by a factor of 4 (i.e. fast interpolation) when a new transition occurred while an interpolation phase is ongoing,

- The crossfade. When a 3- or a 4-sector step occurs, one should not resort to interpolation for 2 main reasons. First, this would slow down the algorithm and result in an annoying delay of the spatialized speech compared to the real location of the speaker. Second, the start and stop filters used in the interpolation would be strongly discrepant, yielding intermediate interpolated spatial filters that do not match any real HRTF. In this case, it is preferred to resort to crossfading so as to ensure a very fast, but still smoothed, transition. The current frame is thus filtered twice: once with the previous filter and once with the current one. These 2 frames are then multiplied by a sigmoid function and added together. Thus, the output frame begins with the previous filter and ends with the current one. When a crossfade is requested while an interpolation is ongoing, the last interpolated filter is taken as the start TF of the crossfade (one does not wait the interpolation to stop, and it is accelerated instead). Crossfade is also used when a change of ITD occurs during an interpolation phase, to avoid perceivable artifacts: the resulting frame begins with the previous ITD and ends with the new one.

4.5 Preliminary subjective evaluation by normal-hearing listeners

This last part deals with the first impressions of some users of the prototype that took part to an informal listening test. For details about the algorithm implementation, the reader should refer to Appendix D. Their observations and comments are reported.

4.5.1 Setup

24 NH listeners (8 at EPFL and 12 at Phonak) have tested the prototype. They were asked to wear the hardware, i.e. the 2 HAs (Phonak Naida IX SP) with the wires connected to the BWU in its box, as shown on Figure 4.19. The wires were stuck against the skin, so as to limit their radiation effect. All the HA features (compression, noise reduction...) were disabled. A led displayed different colors depending on the determined sector, and allowed to follow the good performance of the localization, without hearing the spatial rendering.

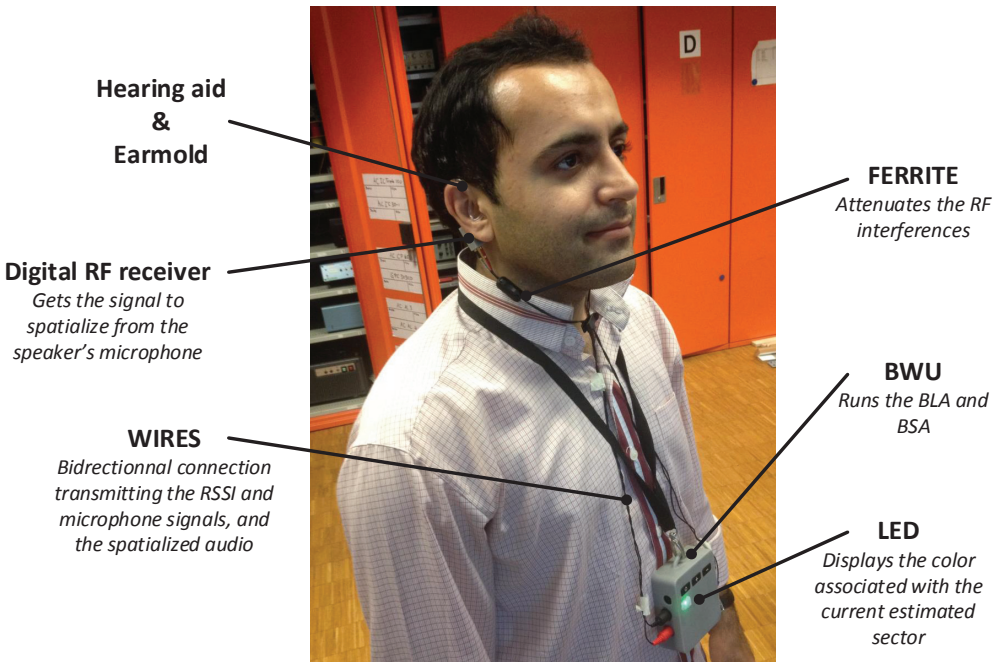


Figure 4.19 – Example of a tested subject wearing the BWU.

4 different subjects (1 female, 3 males) were alternatively speaking. They were wearing the emitter (Phonak Roger inspiro) connected to the BWU. The speaker was moving around the listener at a distance of 3 to 5 meters while reading a text. Their voice was captured by the remote microphone and spatialized in the HAs of the listener. The resolution of the algorithm was fixed and set to 5 sectors.

The experiment took place in 4 different acoustic environments, depending on the tested listener: a listening room, a classroom, a meeting room or an auditorium. After having freely used the prototype for 5 to 10 minutes, the listeners were asked to give their impression, and all the formulated remarks were collected. It was decided not to indicate prior specific features to take care of (such as the reactivity of the algorithm, the externalization, the realism of the rendering...) in order to avoid influencing their attention. Note that the listeners were free to move their head, and could stand up or sit down. This was to ensure realistic test conditions.

4.5.2 Observations

Some observations have been made while conducting the test. They are related to the behavior of the algorithm in real-world environments [42]:

- When the listener is close to a window (10 to 50 cm), the amount of error can significantly increase, which is due to reflections,
- The slowness of the algorithm rises with the increasing reverberation of the room. This is because reverberation yields a higher number of adverse frames. Even though they are mainly detected and discarded, the algorithm requires a longer time to get enough information to update,
- The performance of the algorithm seems to be independent of the speaker, that is, female and male voices have led to the same kind of spatial renderings, which is in agreement with the previous reported results.

4.5.3 Comments

Table 4.6 reports the major positive and negative comments expressed by the listeners during the experiment.

Positive comments	**Negative comments**
1. Spatial effect described as "impressive", "exciting", "clear and nice".	1. Localization error: spatialization sound moving while the speaker is immobile.
2. Reactivity sufficiently fast for head movements.	2. Loudness difference depending on the sector.
3. Occasional errors and slowness judged as "not annoying".	3. Occasionally slow: 1 or 2 s to react to a motion and perform the transition.
4. Spatial rendering convenient even when the speaker is behind.	4. Difficulties in perceiving the difference between L1/L2 and R1/R2.

Table 4.6 – Positive and negative comments expressed by the listeners. Taken from [42].

Overall, the final version of the localization and spatialization algorithms is well appreciated by the listeners, who have been enthusiastic about the spatial effect provided by the system. When localization errors or slowness have occurred, they have not perceived it as significantly boring. The question of using generic HRTFs came when a subject mentioned that he perceived the sound behind him when closing the eyes. But the subject was able to adapt his perception when seeing the speaker in front of him. An important thing is that all tested subjects were directly or indirectly involved in the research and development of this functionality. They were experienced with those kinds of evaluations and were aware of the limitations of the current WMS in terms of processing and rendering. For sure, that has been a bias to consider in this basic study.

4.6 Conclusion

The literature dealing with sound spatialization is well established. In particular, binaural spatialization shows to have lots of applications in a broad range of fields. As for the development of the BLA, the technical constraints has guided and led the research. This has been really appreciated in order to avoid loosing time by roaming in any direction. Indeed, a great part of the previously reported techniques to optimize the offline design of spatial filters and the real-time implementation of them have been exploited:

- The minimum-phase representation of the HRIRs. This is an efficient and lossless way of simplifying these IRs. It facilitates the design of low-order digital filters by concentrating all the energy in the first samples. Also, it enables to introduce the ITD via a pure delay (i.e. a sample shift),
- The design of low-order FIR and IIR spatial filters. The first type of filters has been favored because it is unconditionally stable. Another decisive advantage over IIR filters is that FIR filters are quite a bit easier to interpolate in real time. The filter order have been decreased down to 10, which is beneficial for both memory storage and computational cost. The limited bandwidth of HAs has helped to this end as well,
- The implementation of warping, of which the goal is to reduce the order of the filters by focusing the design on the frequency band of greatest interest. However, the resort to warped filters has been given up once FIR filters has been chosen,
- The 5-sector resolution that translates into the storage of only 5 HRTF pairs. Due to the symmetry of those sectors towards zero, only 5 HRTFs are actually required, plus 2 for the case of a 3-sector resolution,
- The integration of a linear HRIR interpolation to avoid storing tens of intermediate spatial filters.

On the other hand, the previous art related to the target of HI subjects as end-users of binaural spatialization is very scarce. This brand new application is in an early stage. Only 3 studies, published in 2010, 2012 and 2014, investigate the effect of an artificial spatial hearing in listeners with a HRL. 2 other articles (in 2012 and 2013) report the integration of spatial filters in HAs. The main conclusions of interest for an application on WMS are the following:

- The use of customized HRTFs is not useful and generic HRTFs should be sufficient for spatialization in the FHP,
- The constant training that will occur via the visual cue will most probably improve the localization performance of HI subjects with impersonalized HRTFs,
- The perception of externalization is not demonstrated in HI listeners, but it worths keeping an investigation in this direction,

- The low frequency sampling of HAs does not prevent from perceiving a spatial effect in NH listeners. There is a great chance that this is also true for HI subjects, who are used to hearing and localizing without any HF content.

The original contribution of this thesis in the field of binaural spatialization is the concept of HRTF magnitude limitation. The underlying idea is that it is prominent to take into account that the spatialization techniques are performed on HAs, with HI listeners. Appendix B.1.2 reports the limited auditory dynamics of the impaired AS, as well as the necessity to implement dynamic compression in the HA processing. The limitation of the spatial filter gain, developed to consider this related issue, has been evaluated on 38 NH listeners. This extensive psychoacoustic study validates the concept.

Once the code has been implemented in the prototype, it has been possible to collect the first impressions and comments of some NH users. However, one has to remind that those tested listeners were all people aware of the goal of the research. Anyhow, they have been enthusiastic, reporting only little concerns about the slowness of the algorithm. The primary expressed drawbacks are actually related to the spatialization process with problems of generic HRTFs and loudness. The behavior of the system in real conditions is consistent with the expectations coming from the offline runs of the BLA.

One of the main limitations of this primary subjective assessment is the fact that it has not been conducted on inexperienced and especially HI subjects. In order to go further in the evaluation of the developed processing, Chapter 5 introduces and reports the results of a clinical trial designed to assess the perception of the spatialization feature on unexperienced NH and HI listeners.

5 Evaluation of binaural spatialization on hearing-impaired subjects

This last chapter deals with the evaluation of a binaural spatialization technique on HI subjects. As already mentioned, this is a brand new topic, and no study about this evaluation has been reported in the literature so far, to the knowledge of the author. The development of BHAs (Chapter 1.3), as well as the growing interest in including a spatialization functionality in WMS (Chapter 1.4), demands such an investigation, so as to estimate what is the benefit that this technology could provide. However, performing research on human beings, especially impaired subjects, is strongly framed by several ethical principles. This does not correspond to a simple psychoacoustic test, rather one has to consider it as a clinical trial.

Part 5.1 first introduces and details the protocol of the clinical experiments. The approval from an ethics committee is compulsory to start the study. The submission of a protocol for a clinical trial must justify the needs for such a research, by deeply reviewing the literature (the reader can find this thorough review in Appendix E). This is to end up with a procedure conformed to the state-of-the-art habits and the respect of the ethical legislation. Part 5.2 reports the results of the 3 tests conducted in this clinical trial, namely a speech intelligibility test, a sound localization test, and a preference-rating test. Those are performed to respectively assess the effect of the spatialization functionality on the understanding of the speech content, on the ability to localize the speaker, and on the preference of subjects, between a diotic or an artificial spatial hearing. The outcomes are discussed in part 5.3. Part 5.4 draws the final conclusions of this chapter.

5.1 Protocol

The clinical trial of this thesis is conducted to answer several questions about the perception of the spatialization functionality by HI listeners, as explained in the introduction of this chapter. The protocol of the study is presented here, describing the involved subjects, hardware and stimuli used, and the setup for the 3 tests evaluating the speech intelligibility, speaker localization, and preference rating. The terminology is as follows: the clinical trial comprises 3 tests. Each test is made of certain experiments (i.e. 2 experiments in the intelligibility test, 4 in

the localization test, and 1 in the preference-rating test).

5.1.1 Subjects

A total of 40 subjects took part in the clinical trial. They were split in 4 different groups:

- 10 young adult NH subjects (NH group) with hearing thresholds lower or equal to 20 dB HL on both ears, between 125 Hz and 8 kHz,
- 10 moderate HI subjects (HI-MOD group) with pure-tone averages (PTAs, see Appendix B.1.1 for definitions)[1] between 41-60 dB HL,
- 10 severe HI subjects (HI-SVR group) with PTAs between 61-80 dB HL,
- 10 profound HI subjects (HI-PFD group) with PTAs greater than 81 dB HL.

Groups	Age (Mean (SD))	PTA better ear	PTA worse ear
NH	21 (2)	1.6 (2.2) dB HL	3.2 (2.3) dB HL
HI-MOD	51 (24)	49.5 (4.1) dB HL	56.1 (4.4) dB HL
HI-SVR	64 (21)	67.8 (4.4) dB HL	72.3 (5.6) dB HL
HI-PFD	38 (18)	98.6 (10.4) dB HL	103.4 (9.9) dB HL

Table 5.1 – Statistics related to the 40 patients, averaged in each group.

These categories are in agreement with the ones defined by the WHO, reported in Appendix B.1.1. The control group of 10 NH subjects serves as a reference of some normal performance in terms of intelligibility and localization. Additionally, the group is expected to be the most sensitive to the spatialization functionality, especially in the preference-rating test.

All participants had to present an otoscopy within normal limits, as checked by an ear inspection at the beginning of the test. Additionally, all HI subjects suffered from a symmetrical HRL that did not differ by more than 20 dB between the left and right PTAs. This criterion is less constraining than the ones reported in the studies reviewed in Appendix E. There are 2 main reasons for that. First, it facilitates the recruitment of a sufficient number of subjects. Second, it allows to end up with more general conclusions than a study involving a very precise kind of HRLs. Note that a subgroup of 11 FM-experienced listeners was present among the 30 HI subjects. 2 of them were from the HI-MOD group, 3 of them from the HI-SVR, and the last 6 patients belonged to the HI-PFD group. Subjects were paid for their participation.

The criteria concerning the health history of the patients were the following:

- No history of chronic or terminal illness, psychiatric disturbance, or senile dementia, as reported by the participant,

[1] All PTAs are computed at the better ear.

- No history of strong tinnitus or hyperacusis (abnormal acuity of hearing) as reported by the participant,
- No history of stroke or cerebral vascular disorder with a paresis or aphasia as reported by the participant,
- No history of epilepsy or other reactions associated with the proximity to a video screen as reported by the participant,
- No visual impairment, after correction with glasses or not as reported by the participant,
- Willing and able to give written informed consent to participate in this investigation.

All the HI participants were patients of the audiologist involved in the clinical trial, and worn bilateral Phonak HAs in their everyday-life.

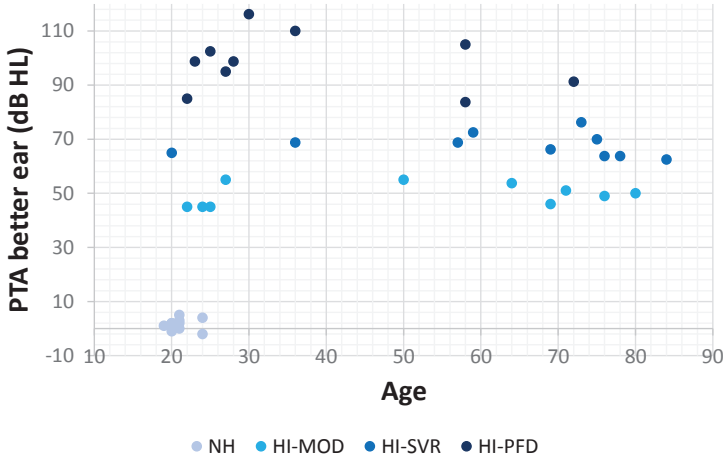


Figure 5.1 – Distribution of the PTA at the better ear as a function of the age of the 40 subjects.

Table 5.1 reports the statistics related to the 4 groups of patient. The displayed PTAs of the best and worst ears, averaged over all patients in each group, show that the HRLs of the patients were symmetrical in all groups. Also, it is noteworthy that the average HRLs in the better ear are well centered in the intervals of each category, so as to diminish the risk of an overlapping effect. Note that the difference of PTAs between the HI-MOD and the HI-SVR groups is on the order of 20 dB HL, while the difference between the HI-SVR and HI-PFD is around 30 dB HL. This is because all the patients suffering from a HRL higher than 81 dB HL are included in the HI-PFD group, would they have a PTA of 85 dB HL or 115 dB HL. Finally, it can be noticed that the patients of the HI-MOD group, and especially of the HI-SVR group, are considerably older than the ones in the HI-PFD group. This is highlighted by Figure 5.1, which represents the PTA at the better ear of the 40 patients as a function of their age. The NH group is clustered within

the 18-25 y.o. interval. Only 2 patients presenting a severe HRL are younger than 40 y.o., and the others are all beyond 55 y.o.. This intimates that the majority of the patients suffering from a severe HRL are actually presbycusic persons. On the contrary, 7 over 10 patients from the HI-PFD are below the age of 40.

Figure 5.2 reveals the origin of the hearing impairment for the 30 HI subjects. 57% of them were born with their HRL, or acquired it during the childhood. Thus, one can consider that they have never benefited from a normal spatial hearing. Conversely, the hearing disorder of 40% of the patients has appeared throughout their life, meaning that these subjects have experienced a normal hearing.

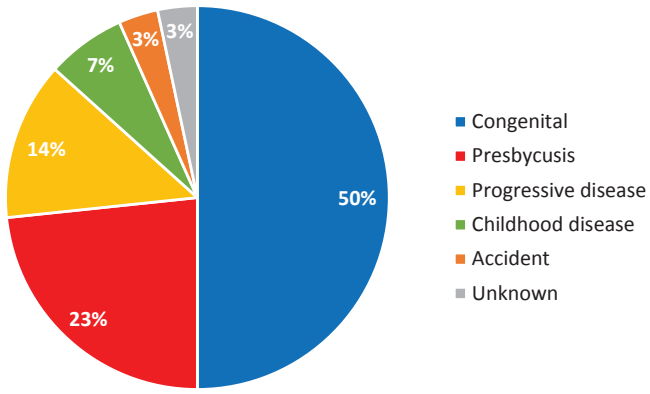


Figure 5.2 – Origin of the HRL of the patients involved in the clinical trial.

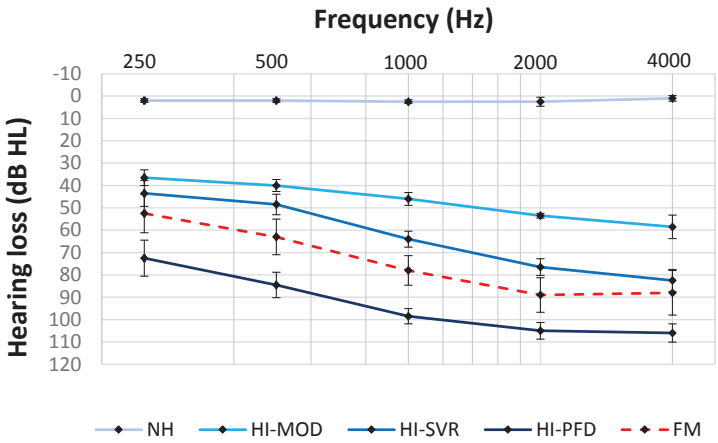


Figure 5.3 – Average audiograms in each category of patients. The average audiogram of the FM-experienced subgroup is in red dashed line.

The average audiograms of the better ear in each group are depicted on Figure 5.3. The mean audiogram of the subgroup of the FM-experienced listeners is displayed as well. Note that all over this chapter, the error bars represent the standard error, that is, the SD divided by the square root of the participant number in each group. Some sloping-HRL with different slopes are shown in all groups. The overlap between the categories is very small, justifying the arrangement of the patients in 3 groups. The FM-experienced subgroup present the HRLs with the strongest variations, because it is composed of patients with various degrees of HRLs, as indicated before.

5.1.2 Pre-test operations

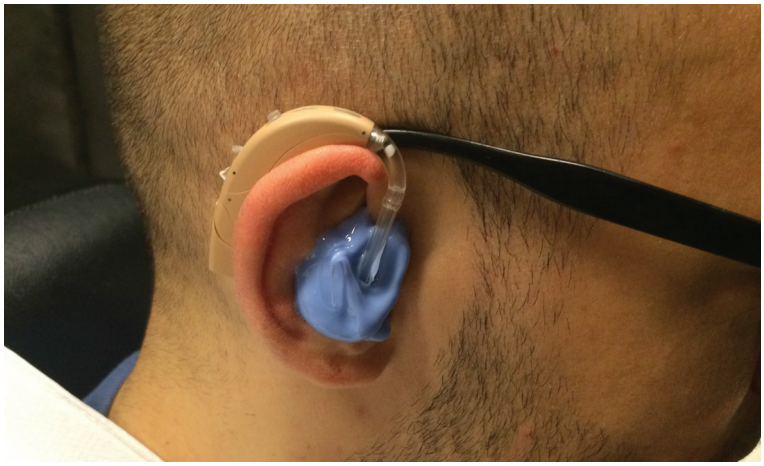


Figure 5.4 – An ear filled with ear impression material.

Prior to the beginning of the test, an otoscopy was performed, and the NH participants went through a pure-tone audiometry over 7 tested frequencies (125, 250, 500, 1000, 2000, 4000, and 8000 kHz), in order to ensure normal hearing. Then, the audiologist made them wear the HAs, and filled their concha with some impression material so as to occlude them, as shown on Figure 5.4. This is to strongly diminish the contribution of the direct sound during the test.

Concerning the HI patients, they had to give their HAs to the audiologist so that he could activate the DAI, and deactivate the unwanted signal processing features. The status of the features in the devices are reported in Table 5.2. For their detailed role, one can refer to Chapter 1.1.2. Those algorithms were switched off because they could interfere with the experiments, they would be useless in the test context, and they could add an undesirable processing delay. On the contrary, the kept features were judged prominent for the HI patients, even though they distort the spatialization rendering *per se*. At the end of the experiment, the audiologist reset the features as they were at the beginning.

The HA of the better ear was submitted to a short calibration, in order to characterize its IN/OUT behavior. This procedure was in agreement with the ANSI S3.22-2003 (paragraph 6.15.1, “Input-output characteristics”) [5], except that the signal used was either a speech

sequence or a speech-shaped noise, instead of a pure-tone signal. This was done so that the reference stimuli was closer to the ones of the study.

Feature name	Action	Status
WDRC	Non-linear amplification	ON
SoundRecover	Non-linear frequency compression	ON
LarsenBlock	Feedback canceller	ON
Directivity	Directional microphone mode	Omni
NoiseBlock	Noise reduction	OFF
WindBlock	Wind-induced noise reduction	OFF
SoundRelax	Limitation of impulsive signals	OFF
Real Ear Sound	Simulation of the pinna filtering lost with BTE models	OFF
SoundFlow	Adaptive program selection depending on the environment	OFF

Table 5.2 – Status of the signal processing features embedded in Phonak HAs.

The knowledge of the IN/OUT characteristics was mandatory to ensure some accurate SNR values during the intelligibility test. Indeed, since the amplification is non-linear, one has to know which gain must be applied to achieve a desired SNR. For the HAs used by the NH group, the characterization was done once in advance. Figure 5.5 shows the dynamic curves for the 4 groups, measured in the 2cc coupler for a speech signal, as a function of the RMS value of the electrical signal input via the DAI. The working level is around 6 dB mV (standard reference level at the input of the DAI). In this area, the amplification provided by the HAs of the NH group is linear. Conversely, if one wants to reach a SNR of 3 dB related to the working level of 114 dB SPL in the HI-PFD group, the gain applied to the masker must be -15 dB. In the NH-SVR group, it is equal to -12 dB with a working level of 98 dB SPL.

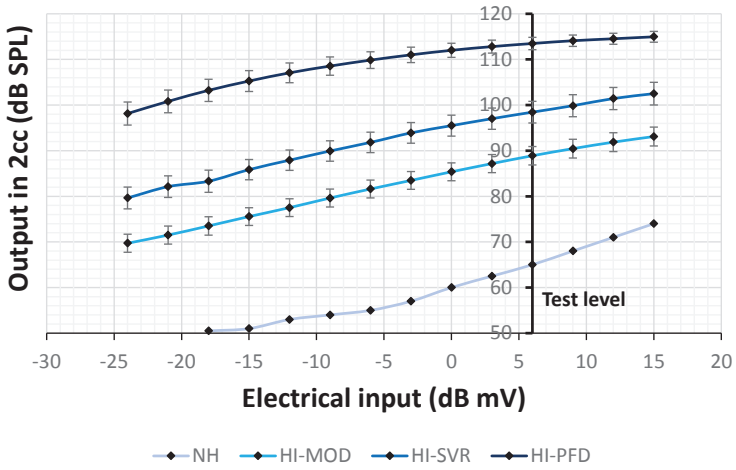


Figure 5.5 – The average IN/OUT characteristics of the subjects' HAs in the different groups.

5.1.3 Stimuli

2 databases of speech content were used: the HINT database, introduced in Appendix E.1, for the intelligibility experiments, and the SUS French database, for the localization test. For the intelligibility experiments, the masking noise was a mixture of 5 uncorrelated speech-shaped noises spatialized in the 5 spatial locations considered in this study (0°, ±30°, ±65°). Each speech-shaped noise had a similar long-term average spectrum as the one of the speech stimuli. The same spatial filters as the ones employed for spatializing the speech signals were used for the masker.

For the preference-rating experiment, the stimuli were some video sequences, recorded with 2 speakers at EPFL, for the specific purpose of the clinical trial.

5.1.4 Hardware

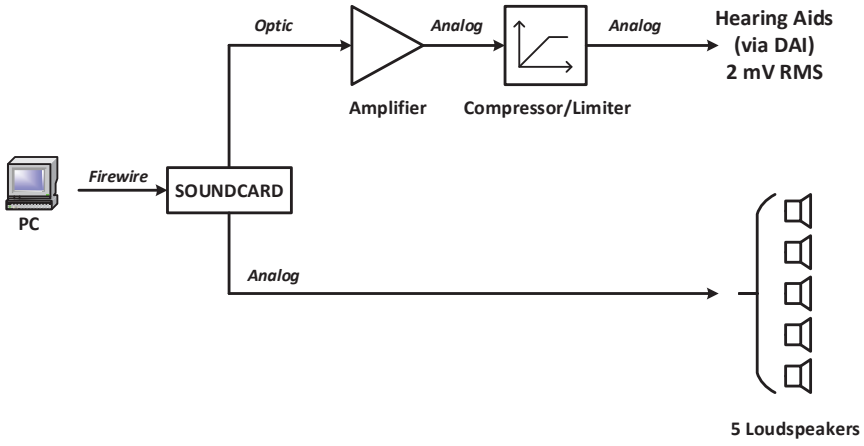


Figure 5.6 – The audio chain mounted for the clinical trial. The nature of the connections between the devices is shown as well. Taken from [46].

The use of the prototype has not been possible because it has no CE certification. Moreover, its behavior between subjects would have been difficult to control. That is why it has been necessary to resort to a different hardware. Figure 5.6 depicts the audio installation, which was composed of the following devices:

- A soundcard MOTU 896 mk3,
- An amplifier Denon AVR 3300, which got the input signals from the soundcard via an optical connection, and reduced the voltage so that it equaled 2 mV RMS,
- A compressor/limiter Samson S-com plus, which prevented the output to exceed the desired level, thus protecting the ears of the subjects in case of an accidental excessive level,

- 5 Tannoy Reveal Active loudspeakers.

A pair of BTE HAs (Phonak Naida IX SP) with standard fittings (matching an audiogram with no HRL) was available for NH subjects. The compatible models of Phonak BTE HAs with the hardware (i.e. incorporating a DAI) are listed in Table 5.3, with the corresponding output power. The HI patients kept up their usual earmolds, and all types of molds and vents have been observed, depending on the degree of HRL (refer to Chapter 1.1.3 for detail). The DAI of the HA was connected to the output of the compressor/limiter via a ground-loop-isolator (Contrik ZNPKL-CHS-CHS), which removed any DC component that could damage the HAs. A chin rest (Orthoptix mentonnière) was used to fix the head and prevent head motion. For the preference-rating test, a large screen was mounted in the test room. A beamer displayed the movies used for the test.

Model	Power microP	Power 13	Power M	Power P	Power SP	Power UP
Ambra	×		×		×	
Audeo	×					
Bolero Q		×		×	×	
Bolero V				×	×	
Cassia	×		×		×	
Certena Art	×		×	×	×	
Dalia	×		×		×	
Exelia	×		×	×	×	
Milo Plus	×				×	×
Naida					×	×
Naida S					×	×
Naida Q					×	×
Sky		×			×	×
Solana	×		×		×	
Versata Art	×		×	×	×	

Table 5.3 – Models of Phonak HAs compatible with the protocol.

5.1.5 Procedure

Intelligibility test

The goal of this test is to establish whether the spatialized speech has a significant effect on the intelligibility, compared to the diotic presentation, in both the FM-only and FM+M modes. In particular, this test must indicate whether the spatialization functionality modifies speech understanding, due to the loss of the binaural summation (Chapter 1.4). Several azimuths were tested to determine if there is an influence of the source position on the speech perception as well. There were 2 experiments, one corresponding to the FM-only mode, and the other to the FM+M mode, as detailed in Table E.4. The procedures are the same for both experiments, except when notified.

The reference level of the speech signal for the NH listeners was set to 65 dB SPL. The gain of the HAs had been adjusted such that an input of 2 mV RMS yielded an output of 65 dB SPL in the 2cc coupler. When it comes to HI subjects, the standard reference level of 2 mV was also input via the DAI. This is assumed to be a comfortable level, as it corresponds to a sound of 70 dB SPL picked up by the HA microphones[2]. A variation of ± 8 dB around the 2 mV level was possible if the subject complained about a too loud or a too soft level. In the FM+M test, the contribution of both sources had to product a level of 65 dB SPL in the ears of the NH listeners. The gain of the loudspeaker was thus set so that a SPL of 67 dB was measured at the HA microphones. The signal passing through the DAI was also reduced by 3 dB. Thus, the addition of both sound sources produced a SPL of 65 dB SPL. The same principle was used for the HI groups.

The listeners sat in the centre of the test room. The speech and noise signals were input via the DAI of the HAs in the FM-only experiment, whereas the speech was played via the DAI and through a loudspeaker at the corresponding location in the FM+M experiment. Note that, in this last case, the noise was rendered only through the DAI. The spatialization was performed in one of the 5 considered directions. In the FM+M experiment, the speech was present in a diotic way via the DAI and played either at -65°, 0° or 65° via the loudspeaker. The sentences were played back at 3 different SNRs. There were 2 sentences per sector and 3 diotic ones, giving a total of 39 sentences. After each sentence, the participant was asked to repeat what was heard. The sentence could not be listened twice. The examiner underlined the words that were correctly understood on an answer sheet. The order of the sentences was the same for all listeners, but the diotic/spatialized conditions were randomized by the test software. The NH and HI subjects were not tested with the same SNRs. The NH subjects experienced some SNRs of -10 dB, -13 dB and -16 dB. For the HI patients, the procedure suggested by Lewis *et al.* [136], reported in Appendix E.1, was adapted to the test. The examiner fixed a certain SNR for the 13 first sentences, after having discussed with the patient about their speech understanding in noisy situations. Then, depending on the results, 2 other SNRs were experienced, such that:

- SNR 1 yielded the better intelligibility score,
- SNR 2 yielded an intermediate intelligibility score,
- SNR 3 yielded the worst intelligibility score.

Table 5.4 indicates the average SNRs experienced by the listeners in the 4 groups. As expected, the stronger the HRL, the higher the required SNRs. Apart from the HI-PFD group, the step between the SNRs 1, 2 and 3 is of about 3 dB. The profound HI patients exhibit the highest SD, which is due to the various PTAs gathered in this group. All SNRs were monitored by varying the noise level, while keeping the speech level fixed, in agreement with the studies reported in Appendix E.1.

[2] For all Phonak HAs, an electrical signal of 2 mV RMS at the DAI and an acoustical signal of around 70 dB SPL at the HA microphones generate the same output SPL.

Groups	SNR 1 (mean (SD))	SNR 1 (mean (SD))	SNR 1 (mean (SD))
NH	-10	-13	-16
HI-MOD	-3.9 (3.2)	-7 (3.4)	-9.7 (3.3)
HI-SVR	-1.3 (2.2)	-4.3 (2.2)	-7.3 (2.2)
HI-PFD	6 (4.9)	4.7 (7.5)	2.3 (8.8)

Table 5.4 – Average SNR experienced by the participants, in the 4 groups.

The test started with a training procedure of 6 test sentences (1 diotic + 1 in each spatialized direction), such that the listeners could get used to the procedure, and hear the various spatial conditions once. The listeners were not aware of the real beginning of the test after this training period.

To ensure an equal loudness in the different spatial directions, the following calibration step had been conducted. The long-term RMS value of a speech-spectrum noise had been measured in the 2cc coupler for the diotic and 0°-spatialization cases, when played through the DAI of a HA. The levels had been adjusted until the corresponding loudnesses were the same. The other spatialized directions were supposed to yield the same binaural SPL as the one at 0°. An example is depicted on Figure 5.7. The spectrum of a certain sequence of speech is depicted in red, and present a loudness of 65.8 phons (see Appendix A.1.2 for definitions), as calculated with the loudness Matlab toolbox developed by the Genesis company [79], from the methods proposed by Glasberg and Moore [82]. After filtering with the spatial filter at 0°, the loudness equals 68.4 dB, which would artificially increase the SNR of the processed sentences. Therefore, a gain is applied, so as to recover the original loudness.

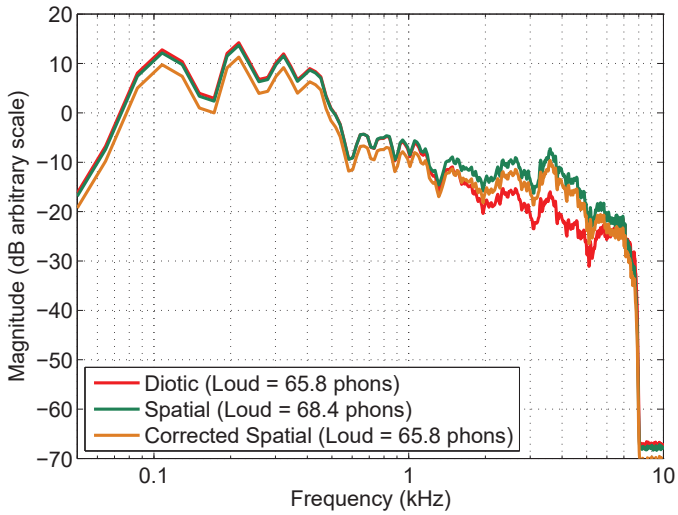


Figure 5.7 – Spectrum of a diotic speech sequence (red), of the same sequence spatialized at 0°, before (green) and after (orange) loudness equalization.

Localization test

The objective of the localization test is to evaluate the effect of the binaural spatialization on the localization performance of NH and HI listeners. It is composed of 4 experiments, as described in Table E.7. The first determines the unaided performance of the subjects. The second investigates the localization abilities obtained with HAs. The third is related to the implementation of the BSA in the FM-only mode, and the fourth is the same in the FM+M mode.

In the localization test, the subjects sat in the centre of the room. They were asked to fix the front, and their head was immobilized by the chin rest. The listeners could not see the different loudspeakers, which were hidden by a black curtain, as shown on Figure 5.8. 9 numbers from -4 to 4 were displayed on the curtain, corresponding to azimuths at 0°, ±15°, ±30°, ±45°and ±65°. This procedure was similar in the 4 experiments.

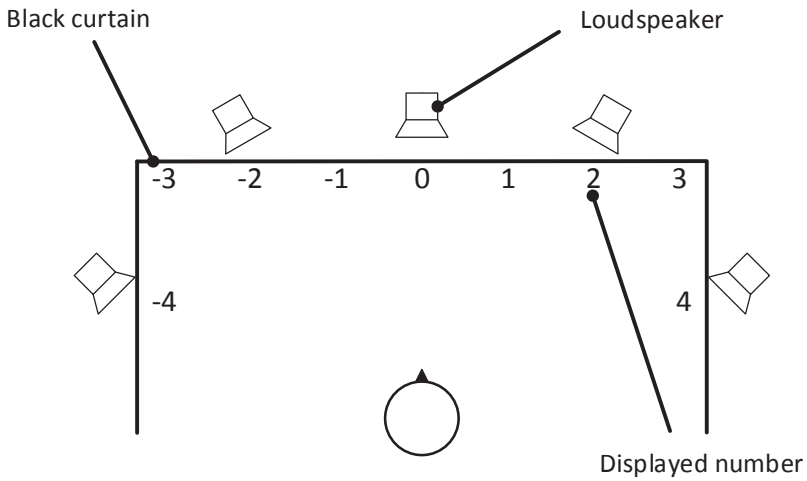


Figure 5.8 – Setup for the localization test. Taken from [46].

In these experiments, the sentences were spatialized in one of the 5 sectors. 3 sentences are played in each direction, resulting in a total of 15 sentences. In the FM+M mode, the stimuli were rendered simultaneously via the DAI and through the loudspeaker corresponding to the current spatialized direction.

The subjects were played successive sentences from the SUS database. In the unaided experiment, the NH participants were played the sentences at 65 dB SPL in all directions. For the HI patients, the output level was individually set to be comfortable for each of them. The average SPL in the 3 groups equaled 77.7 dB SPL, 83.3 dB SPL and 87.8 dB SPL respectively. A maximum level of 98 SPL has been delivered for a patient. A roving level of 6 dB among stimuli was implemented (see Appendix E.2), i.e. the stimuli were played at any random level between -3 dB and +3 dB relative to the static SPL. In the aided experiment, the NH listeners

experienced a level of 65 dB SPL in the ear canal for all directions. For the 2 last experiments, an input at 2 mV RMS via the DAI was used, with the possibility to adjust the level between ±8 dB for the HI patients.

The listeners were played the sentences, one after the other. After each sentence they were asked to indicate the perceived location of the sound source by reporting the number corresponding to the incidence direction. They were made aware that all the available locations may not be played. They could also answer that they perceived the sound from none of those directions. The sentences could not be repeated. The order of the sentences was the same for all listeners, but the spatialized directions were randomized by the test software. In the FM-only and FM+M experiments, the sentences were spatialized in one of the 5 sectors. In the FM+M mode, the stimuli were rendered simultaneously via the DAI and through the loudspeaker corresponding to the current spatialized direction. 3 sentences were played in each direction, resulting in a total of 15 sentences.

All the experiments started with a training procedure of 5 test sentences in each spatialized direction, so that listeners could get used to the procedure and hear the various spatial conditions once. Listeners were not aware of the real beginning of the test after this training period.

Preference-rating test

The goal of this third and last test is to collect the preference of the subjects between the current diotic rendering and the one which is suggested with the BSA. It rests upon some video sequences, as details in what follows.

The listeners sat in the centre of the room. They were wearing their HAs, which were connected to the soundcard through the DAI (FM-only mode). The subjects were asked to stare at the front, and their head was immobilized with the chin rest. They were facing a large screen, and a beamer displayed several movies that show either one dynamic speaker (moving from the right to the left) or 2 static speaker(s) (at the left and right side of a table in a classroom). This last case is illustrated on Figure 5.9, which is a picture taken during a test session. The sound rendered through the HAs was alternatively diotic or spatialized. 3 qualities of spatialization were tested:

- Good spatialization: The spatialization followed the speaker's position in real time (e.g. if the speaker moved from 30° to 65°, the speech signal was instantaneously spatialized from the sector L1 to the sector L2,
- Delayed spatialization: The spatialization followed the speaker's position with a delay of 1.5 second, corresponding to the average delay of the prototype (localization plus spatialization delays),
- Wrong spatialization: the spatialization did not match the speaker's position (e.g. if the speaker was on the right, the spatial filters from the left side were applied).

Figure 5.9 – Picture of the setup for the preference-rating test.

All the audio stimuli were prepared in advance. They were played at 65 dB SPL for the NH subjects, and at the comfortable level determined in the intelligibility test for the HI subjects. A total of 6 scenarios was considered, as reported in Table 5.5. Every scenario was made of 2 different audio versions. Each audiovisual stimuli lasted around 10 seconds.

Scenario	Number of speakers	Motion	Spatialization quality
1	2	Static	Good
2	2	Static	Delayed
3	2	Static	Wrong
4	1	Dynamic	Good
5	1	Dynamic	Delayed
6	1	Dynamic	Wrong

Table 5.5 – The 6 different scenario displayed in the preference-rating test.

The speakers were reading a simple text (sentences from the intelligibility test). Their voice was clear and pleasant, and the pronunciation was well articulated and neutral. The films have been captured using a camcorder located at the virtual place of the listener in a classroom at EPFL. The sound has been recorded using 2 Phonak Roger inspiro devices.

Before starting the test, the participants were made aware of the task they had to perform and

the features they had to focus on. They were asked to assess 5 sound attributes, by answering some questions. These attributes and related questions were the following:

1. Intelligibility: Which stimulus provides the easier-to-understand speech?

2. Naturalness: Which stimulus sounds the most natural for you?

3. Pleasantness: Which stimulus is the most pleasant to hear?

4. Immersion: With which stimulus do you feel better immersed in the room?

5. Overall preference: Which stimulus do you prefer?

The attributes were clearly explained with both a written and an oral explanations. All subjects were submitted to the 6 scenarios in the same order. They were played each scenario, once with the version 1 and once with the version 2. The allocation of the 2 renderings (diotic or spatialized) over the 2 versions was randomized for every scenario. Then, the first version was looped and the listeners could change the version right during the course of the movie by pressing a button. After that, the subjects filled the form for the current scenario, answering all questions by choosing one of the 5-level answer ("Version 1 quite more than Version 2", "Version 1 a bit more than Version 2", "No preference", "Version 2 a bit more than Version 1", "Version 2 a quite more than Version 1"), inspired by the ITU-T P.800 recommendation [106]. The listeners were free to experience the stimuli in both versions as much as they wanted, and they could pause and play the movie when it was desired. The audiologist assisted each participant to ensure that the attributes were well understood, and helped the elderly with the keyboard manipulation. When all the questions were answered, the listeners could switch to the next scenario, until the 6 ones were covered.

Table 5.6 provides a summary of the the listening conditions and stimuli used in the 3 tests previously described.

Test	**Listening condition**	**Stimulus**
Intelligibility	FM-only FM+M	HINT meaningful speech database
Localization	Unaided Aided FM-only FM+M	SUS meaningless speech database
Preference	FM-only	Audiovisual dedicated speech recordings

Table 5.6 – Summary of the listening conditions and stimuli used in the 3 tests of the clinical trial.

5.1.6 Ethical consideration

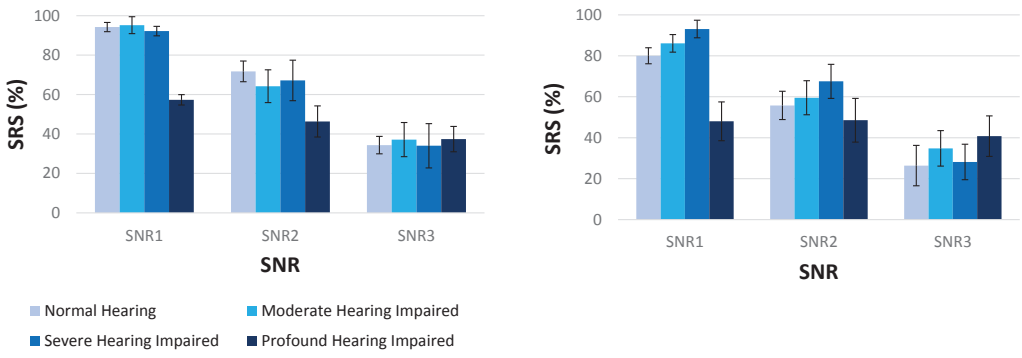
This clinical trial has been evaluated and validated by the Swiss Ethics Committee on research involving humans (swissethics [225]) of the Canton de Vaud (CER-VD [34]) . The protocol is in agreement with the ethical principle of the Declaration of Helsinki of the World Medical Association [8], with the Swiss constitutional article related to the research on human beings [222], with the Swiss law related to the research on human beings [223], and with the order on the clinical trials conducted in Switzerland [224]. This clinical trial, sponsored by Sonova AG, is registered with the identifier NCT02693704 on the website www.clinicaltrials.gov [47] of the US National Institutes of Health, which gathers the information about studies in 193 countries.

5.2 Results

This third part presents the outcomes of the clinical trial, over the 3 tests. Graphs, as well as statistical tests are reported here. The type-I error is fixed at 5% ($\alpha = 0.05$). Each evaluation (intelligibility, localization and preference) is investigated considering the impact of several factors.

5.2.1 Intelligibility test

Intelligibility and SNR



(a) (b)

Figure 5.10 – SRS as a function of the SNR for the different groups in the FM-only mode (A) and in the FM+M mode (B).

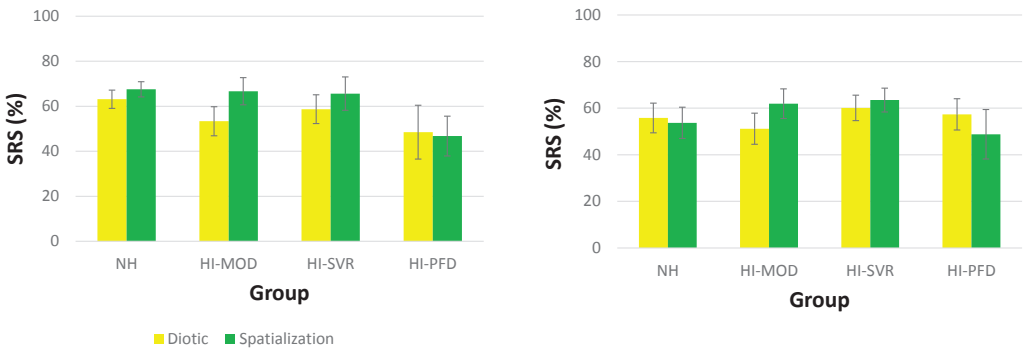
The results of the intelligibility test (expressed in terms of SRS) as a function of the 3 tested SNRs are shown on Figure 5.10, in the FM-only mode (Figure 5.10A) and in the FM+M mode (Figure 5.10B). One has to recall that the SNRs were not the same in the 4 groups, as described in Table 5.4. Therefore, it makes no sense to compare the different SRSs between groups. The outcomes from a 2-way repeated measures ANOVA investigating the influence of the SNR and

rendering mode are reported in Table 5.7. This table reveals that there is a significant effect of the SNR on the speech understanding, i.e. the speech intelligibility decreases when the SNR diminishes for all groups, as shown on Figure 5.10. The test fails to show any statistical influence of the rendering mode. Finally, there is no interaction effect between both factors in the 4 groups (not reported here).

Group	Factor	d.f. 1	d.f. 2	F	p
NH	SNR	2	18	70.298	<0.001
	Mode	1	9	2.357	0.159
HI-MOD	SNR	2	18	28.618	<0.001
	Mode	1	9	1.305	0.283
HI-SVR	SNR	2	18	27.376	<0.001
	Mode	1	9	0.432	0.528
HI-PFD	SNR	2	18	4.516	0.04
	Mode	1	9	1.441	0.284

Table 5.7 – Results of a 2-way repeated measures ANOVA, showing the effect of the SNR and mode on the speech intelligibility for the 4 groups. The significant effects are in red ($\alpha = 0.05$).

Intelligibility and rendering



(a) (b)

Figure 5.11 – SRS, averaged over all SNRs, for the 4 groups, with the diotic (yellow) and spatialized (green) renderings, in the FM-only mode (A) and in the FM+M mode (B).

Figure 5.11 represents the SRS obtained with the diotic (yellow) and spatialized (green) renderings in the FM-only (Figure 5.11A) and the FM+M (Figure 5.11B) modes . The SRS is presented in each group and averaged over all SNRs. In the FM mode, the graph suggests an improvement of the intelligibility with the spatialized rendering in the NH, HI-MOD and HI-SVR groups, while the HI-PFD group seems to show no difference. In the FM+M mode, one can suspect an enhancement of the speech understanding for the moderate HI subjects. On the contrary, a

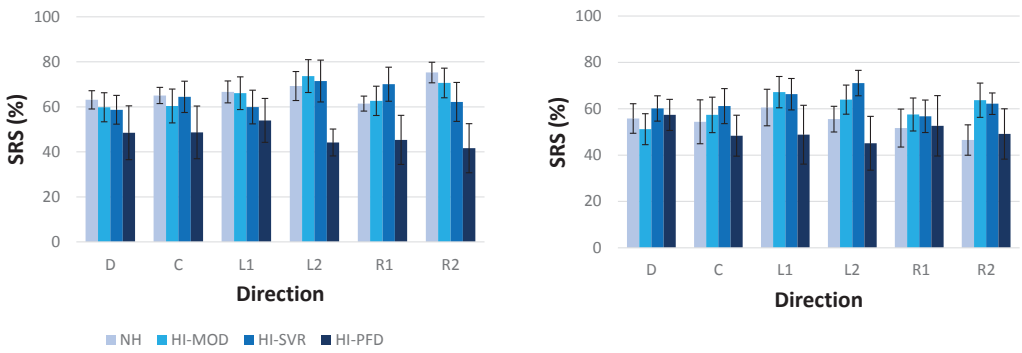
diminution of the intelligibility might occur in the HI-PFD group when the spatialization is applied.

Table 5.8 displays the results from the one-tailed paired-sample t-tests performed in the NH, HI-MOD and HI-SVR groups, in both modes. A one-tailed test is chosen because one wants to test the alternative hypothesis that the spatialization feature improves the speech intelligibility. In the FM-only mode, the analysis finds a significant enhancement of the understanding performance for the moderate and severe HI listeners, while the alternative hypothesis is rejected in the NH group. In the FM+M mode, a significant effect is only present in the HI-MOD group. For the profound HI subjects, the one-tailed tests is performed in the oppositive direction, assuming that the spatialized rendering decreases the intelligibility performance. The tests fails to show any statistical effect in both modes (FM-only: $t(5) = 0.386$, $p = 0.358$, FM+M: $t(5) = 1.142$, $p = 0.153$).

	FM-only		**FM+M**	
Groups	t	p	t	p
NH	-0.66	0.2045	0.486	0.320
HI-MOD	-3.307	0.005	-3.049	0.014
HI-SVR	-4.469	0.001	-0.774	0.230

Table 5.8 – Results of the paired-sample t-tests performed to compare the effect of the diotic or spatialized rendering on the speech perception, in both modes. The significant effects are in red ($\alpha = 0.05$).

Intelligibility and DOA



(a) (b)

Figure 5.12 – SRS, averaged over all SNRs, as a function of the DOA, for the 4 groups, in the FM-only (A) and the FM+M (B) modes. “D” stands for the diotic rendering.

Finally, the influence of the DOA of the speech signal on the SRS is detailed. Figure 5.12 depicts the different SRSs in the 5 sectors and in the diotic rendering. A sequence of one-way

repeated measures ANOVAs have been performed to look for some significant effects of the direction in the results for the 4 groups. In the FM-only mode, a statistical influence is found for the HI-SVR group ($F_{5,45} = 2.998$, $p = 0.020$). A Bonferroni post-hoc test has revealed that the intelligibility is significantly higher in the diotic rendering rather than in the spatialization processed in the sector R1. In the FM+M mode, there is a statistical effect with the moderate HI listeners ($F_{5,45} = 2.579$, $p = 0.039$). However, no effect remained after applying a Bonferroni correction for multiple comparisons.

5.2.2 Localization test

Localization and configuration

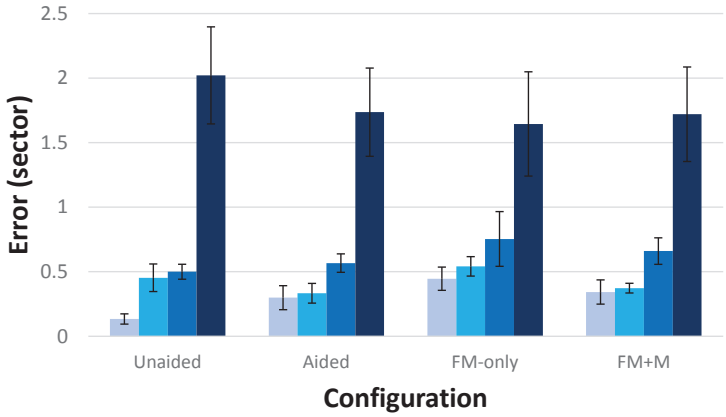


Figure 5.13 – Localization error in the different experiments, for the 4 groups.

The localization error in the 4 configurations is presented on Figure 5.13. Note that this localization error is expressed in sector. Indeed, it makes no sense to report the error in degree, because of the coarse resolution and the fact that the sectors do not present the same angular span. Considering the NH group, Figure 5.13 suggests that there exist some significant differences between the 4 configurations. This is confirmed by a one-way repeated measures ANOVA ($F_{3,27} = 3.837$, $p < 0.05$). A one-tailed Bonferroni post-hoc test indicates that there is a degradation of the localization performance between the unaided and FM-only configurations ($p = 0.036$) . For the 3 HI groups, no significant difference is found between the configurations (HI-MOD: $F_{3,27} = 0.960$, $p = 0.426$, HI-SVR: $F_{3,12.505} = 0.951$, $p = 0.380$, HI-PFD: $F_{3,24} = 0.510$, $p = 0.679$). Note that a Greehouse-Geisser correction has been applied for the HI-SVR case, as the data did not fulfill the sphericity assumption.

Localization and DOA

After the computation of several one-way repeated measures ANOVAs, some significant effects of the sector of incidence on the localization performance arise. This is the case for the NH subjects in the aided ($F_{2.314,20.822} = 4.768$, $p = 0.016$, (Greenhouse-Geisser correction)) and FM-only ($F_{4,36} = 4.454$, $p = 0.012$) configurations. In the HI-MOD group, the results depend on the DOA in all configurations (Unaided: $F_{4,36} = 2.729$, $p = 0.044$, Aided: $F_{4,36} = 10.248$, $p < 0.001$, FM-only: $F_{4,36} = 2.853$, $p = 0.038$, FM+M: $F_{4,36} = 8.847$, $p < 0.001$). Finally, the severe HI subjects also show some statistical effects in the unaided ($F_{4,36} = 4.124$, $p = 0.022$), aided ($F_{4,36} = 3.462$, $p = 0.017$) and FM+M ($F_{4,36} = 4.210$, $p = 0.007$) configurations. Nevertheless, only few effects remain after the computation of Bonferroni post-hoc tests, as indicated in Table 5.9.

Group	**Configuration**	**Significant effect**	p
NH	Aided	Localization better in R2 than L1	0.020
HI-MOD	Aided	Localization better in C than L1	0.004
		Localization better in L2 than L1	0.010
		Localization better in R2 than L1	0.015
	FM-only	Localization better in C than L1	0.040
	FM+M	Localization better in C than L1	0.016
		Localization better in L2 than L1	0.049
		Localization better in R2 than L1	0.009

Table 5.9 – Results of the Bonferroni post-hoc tests reporting a significant effect of the DOA on the localization performance ($\alpha = 0.05$).

Interestingly, Table 5.9 suggests that the intermediate sector L1 provides higher localization errors compared to both the central and extreme sectors. Figure 5.14 depicts the influence of the sector type (i.e. central (CTR), intermediate (INT) and extreme (EXT)) on the localization accuracy. One may suspect some significant differences between the 3 types of sectors. In particular, it seems that the intermediate sectors lead to worse performance compared to the central and extreme ones. Table 5.10 reports a sequence of one-way repeated measures ANOVAs with Bonferroni correction for multiple comparisons applied, which confirms that the localization error is often statistically worse in the intermediate sectors. Also, it appears sometimes that the performance is better in the central sector than in the extreme ones.

The results reported here are not the consequence from a bias in the setup, as evidenced by Figure 5.15. In fact, the sectors L1 and R1, as well as the sectors L2 and R2, yield the same amount of error across all subjects.

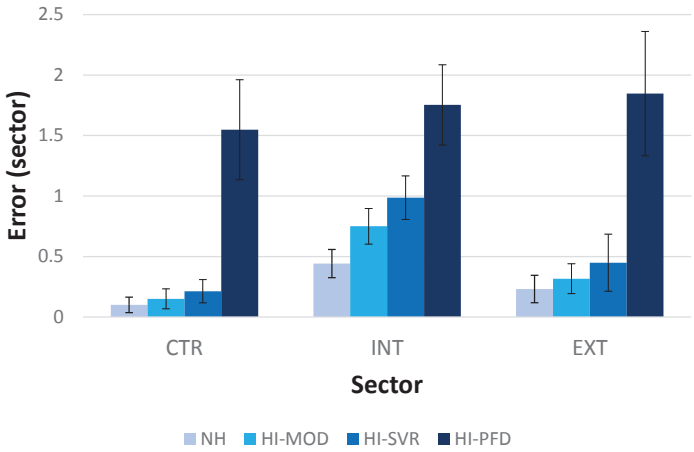


Figure 5.14 – Localization error in the central sector (CTR), in the intermediate sectors (INT) and in the extreme sectors (EXT), for the 4 groups.

Group	Configuration	Effect	p
NH	Aided	Localization better in CTR than INT	0.023
HI-MOD	Unaided	Localization better in CTR than INT	0.047
		Localization better in CTR than EXT	0.016
	Aided	Localization better in CTR than INT	0.001
		Localization better in EXT than INT	0.013
	FM-only	Localization better in CTR than INT	0.016
	FM+M	Localization better in CTR than INT	0.010
		Localization better in EXT than INT	0.010
HI-SVR	FM-only	Localization better in CTR than INT	0.001
	FM+M	Localization better in EXT than INT	0.028
HI-PFD	FM+M	Localization better in CTR than INT	0.045

Table 5.10 – Results of the one-way repeated measures ANOVAs with Bonferroni correction for multiple comparisons, reporting a significant effect of the sector type on the localization performance ($\alpha = 0.05$).

Localization and group

Contrary to the intelligibility test, it is possible to directly compare the performance between the different groups. To this end, several one-way between-subjects ANOVAs have been conducted. First, the HI-PFD group is not considered in the analysis, as it yields a violation of the homogeneity of variance assumption. The null hypothesis states that there is no significant difference of localization performance between the 3 other groups, while the alternative hypothesis claims that there is a significant degradation when the HRL increases. Statistical effects have been found in the unaided ($F_{2,27} = 7.453$, $p = 0.002$), aided ($F_{2,27} = 2.729$, $p = 0.037$)

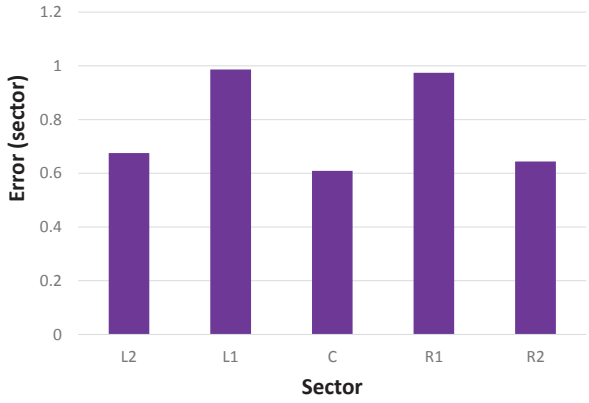


Figure 5.15 – Localization error as a function of the sectors, averaged over the 4 groups.

and FM+M ($F_{2,27} = 4.664$, $p = 0.009$) configurations. Tukey's one-tailed post-hoc tests show a significant increase of the localization error between the NH and HI-MOD groups ($p = 0.006$) and between the NH and HI-SVR groups ($p = 0.002$) in the unaided configuration. Then only statistical differences between the NH and HI-SVR groups have been observed in the aided ($p = 0.031$) and FM+M ($p = 0.013$). Comparing the 2 HI groups, a significant degradation of the performance arises between the moderate and severe HI subjects in the FM+M configuration ($p = 0.024$). Note that no significant difference has been found between the 3 groups in the FM-only configuration.

When it comes to the results of the HI-PFD, it is required to resort to another procedure, according to the previously mentioned reason. A one-way between subjects ANOVA including a Brown-Forsythe correction for unequal variances has shown a significant degradation of the localization performance between the severe and the profound HI subjects, in all configurations (Unaided: $F_{1,8.307} = 14.457$, $p = 0.003$, Aided: $F_{1,9.707} = 11.257$, $p = 0.004$, FM-only: $F_{1,13.205} = 3.9$, $p = 0.035$, FM+M: $F_{1,10.267} = 7.830$, $p = 0.009$).

5.2.3 Preference-rating test

Overall results

The preference ratings of the 4 groups (over the rows), for the 3 spatialization qualities (over the columns) and for the 5 sound attributes are depicted on Figure 5.16. The mention of "Spatialization ++" (dark green) stands for a great preference for the spatialization rendering, and the "Spatialization +" (light green) denotes a little preference for it. This also holds for the diotic rendering, with the yellow and light orange colors respectively. The areas in blue represent the proportion of subjects having no preference between both renderings.

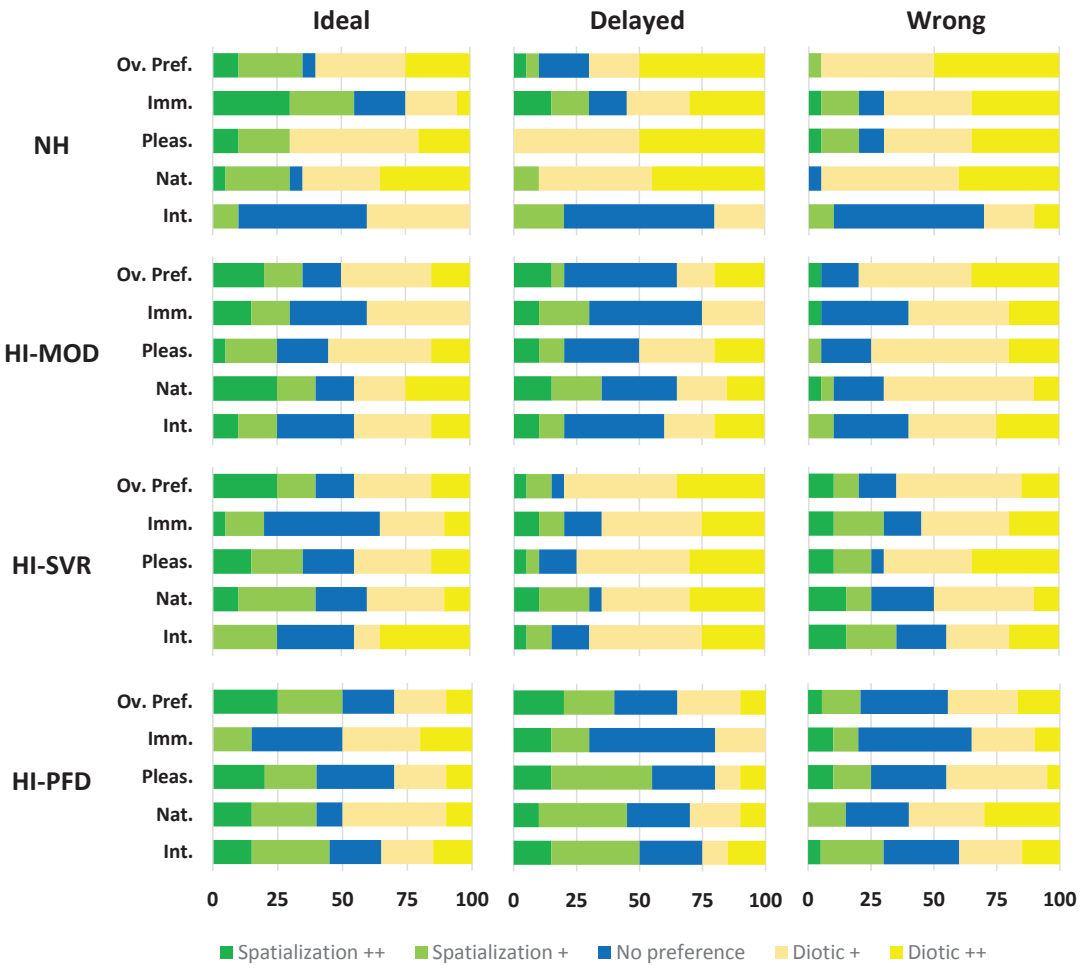


Figure 5.16 – Results of the preference-rating test. The columns are for the 3 qualities of the spatialization (ideal, delayed, wrong), and the rows show the preferences in each group.

Focus on the FM-experienced subjects

Figure 5.17 displays the comparison between the FM-experienced subgroup and the other HI subjects for the ratings of the 3 qualities of spatialization (over the columns). Finally, Figure 5.18 reports the specific cases of the "Naturalness" and "Overall preference" attributes, rated by the FM-experienced patients in the ideal static scenario (2 speakers on a table). A chi-square test of independence shows that there exists a significant dependence between the 2 subgroups on the overall preference in the ideal static scenario ($X(2, N = 30) = 6.489, p = 0.039$). This indicates that the FM-experienced patients do significantly prefer the spatialization feature than the non FM-experienced ones in this scenario.

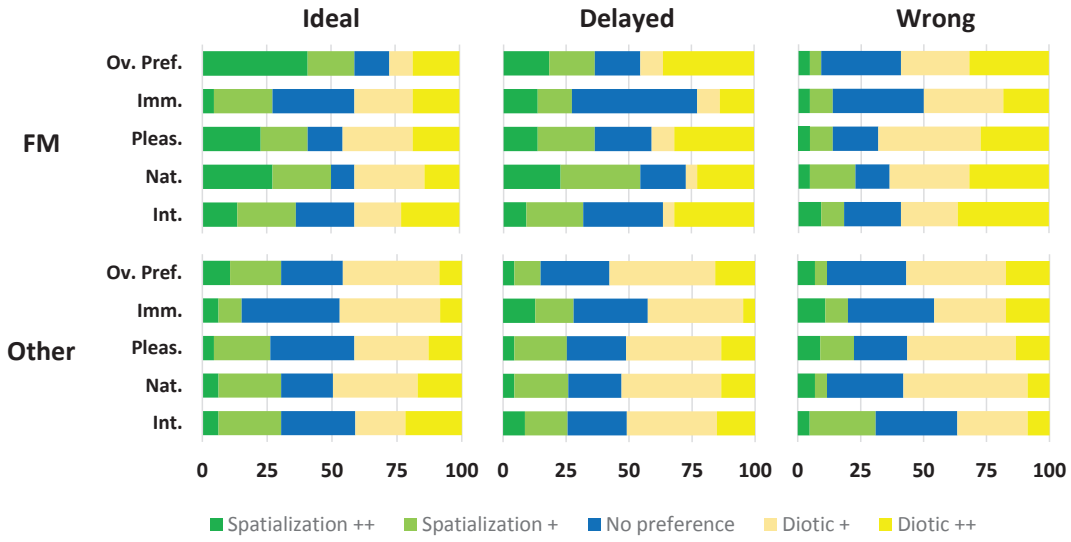


Figure 5.17 – Results of the preference-rating test for comparing the results from the FM-experienced subgroup and all the other HI subjects. The columns are for the 3 qualities of the spatialization (ideal, delayed, wrong), and the rows show the preferences in each group.

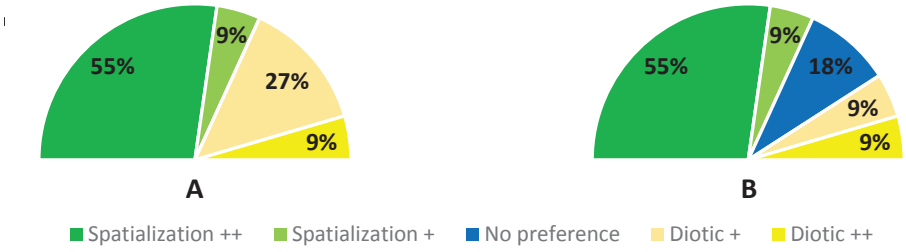


Figure 5.18 – Results of the attributes “Naturalness” (A) and “Overall preference” (B) for the static scenario with the ideal spatialization in the FM-experienced subgroup.

5.3 Discussion

After having reported the detailed results of the clinical trial, one must now analyze and interpret them. The objective is to come up with some reliable conclusions that help understand the effect of the binaural spatialization processing on HI subjects.

5.3.1 Intelligibility test

Intelligibility and SNR

The adapted procedure from Picou *et al.* [187] has yielded an efficient and powerful way of conducting the intelligibility test, finding the adequate SNRs for every subject. The decrease of the SNR degrades the speech understanding for all groups, as could be expected. It must

be mentioned that only 6 profound HI subjects managed to pass the test, while the 4 others did not understand anything, even when no noise was played. Besides, numerous patients, especially in the HI-SVR and in the HI-PFD groups, have indicated that they resort to lip reading in their daily life, whereas the test did not provide this cue. Listeners have reported other remarks: they said that they could hear the speech but not understand its content (a well-known comment, to be linked to their distorted AS, see Appendix B.1.2), that the procedure required much concentration, and that the cognition really helped guess some missing words, like in their real life. This is in agreement with what is evoked in Appendix A.2.2. The choice of a meaningful database is then legitimate. Considering the listening mode, the statistical tests fail to show any significant effect on the intelligibility performance. Note that this test was not designed to estimate the FM advantage (Chapter 1.4.1), as the level of the FM and the HA microphone renderings were fixed.

The participants were not able to benefit from the masking release (Appendices A.1.3 and A.2.2) because the masker was continuous. On the other hand, no temporal masking (Appendix A.1.4) between the noise and speech could occur, since the voice always started 1 second after the onset of the noise. The detailed results among the listeners show that the NH and HI subjects identify the vowels with the (approximately) same performance, but that the HI listeners experience quite a bit more difficulties with the consonants. This exhibits the reduced frequency resolution of the impaired AS, reported in Appendix B.1.1.

Intelligibility and rendering

One of the main result of this evaluation is that the spatialization functionality significantly improves the speech intelligibility of the patients suffering from a moderate hearing impairment in both the FM-only and FM+M modes by an average amount of 9%. Over the 10 subjects, 9 experienced an increase of the intelligibility between 2% and 17%, while only 1 subject presented a marginal loss of 0.9%. The same conclusion can be drawn for the severe HI subjects in the FM-only mode, with a mean intelligibility gain of 7% (min = 0.5%, max = 18%). This is a prominent and appreciated outcome, as it means that the artificial introduction of an ipsilateral and a contralateral ear does enhance the understanding of the speech, despite the partial loss of binaural summation. The loudness compensation between the diotic and spatialized rendering ensures that this observation is not the consequence of some higher SNRs with the spatialization processing. Although it has not been assessed, it is likely that the procedure of HRTF limitation (Chapter 4.3.2) contributes to that result. The depicted results suggest that the intelligibility is preserved for both the NH and HI patients with a profound HRL.

The impact of the signal processing features embedded in the HAs has not been investigated. The choice has been made to keep the WDRC and frequency compression, since it is known that the HI aided subjects take advantage of them to better understand the speech content (Chapter 1.2.2). On the other hand, these features distort the spatial rendering *per se*. Conversely, the noise reduction and microphone directivity has been deactivated. The first provide

no improvement of the SNR, and the second is not available in the WMS. As indicated in Chapter 1.3, the increasing development of some binaural algorithms that preserve the spatial cues leads to some improvement in terms of intelligibility. The present processing goes toward this direction.

Intelligibility and DOA

Numerous subjects have reported that they found the perception of the speech content easier when the voice was coming from the sides. Also, several patients indicated that they perceived a difference of loudness between the sectors, especially in the HFs. However, the statistical analysis failed to show any significant effect of the DOA on the intelligibility, with the only exception of the severe HI patients in the sector R1. The subjects' comment might be related to the fact that the masker was at once spatialized and diotic. Indeed, 5 different uncorrelated noises were spatialized in the 5 sectors and added. Thus, the masker was perceived wider than would have been a unique diotic noise. However, the summation of the 5 noises bring an almost identical sound rendered in both HAs, which is close to a diotic rendering. As a summary, one can say that the noise was diotic but spread out throughout the FHP (i.e. no perception of a compact auditory event in the centre of the head). Note that a spatial separation between the noise and speech would have certainly resulted in a greater improvement of the intelligibility performance from the algorithm, thanks to the spatial release from masking (SRM, see Appendix A.2.2), but this was not the goal of the actual test.

5.3.2 Localization test

Localization and configuration

This study shows that the process of binaural spatialization based on generic spatial filters degrades the localization performance of NH subjects (+20% of localization error). This is in contradiction with what was reported by e.g. Begault *et al.* [14], Drullman and Bronkhorst [61], or Wenzel *et al.* [243] (see Chapter 4.1.2 for the complete review). However, one has to keep in mind that, in the framework of this experiment, the HRTFs are strongly distorted, due to the approximation with low-order filters as well as the resort to amplitude- and band-limited HRTFs. It is likely that a long-term training with these spatial filters would result in some better performance of localization, as shown in the recent studies of Madjak *et al.* [147] and Mendonça *et al.* [167]. Note that this degradation is not significant when the direct sound is available, i.e. in the aided and FM+M configurations.

Concerning the HI subjects, the statistical analysis failed to show such a detrimental effect, whatever the degree of HRL. This somehow suggests that the use of such corrupted HRTFs is not problematic for subjects suffering from HRL. A decrease of the localization error is noticed in the moderate (-27%) and profound (-14%) HI listeners when wearing their HA

compared to the unaided condition, while the patients with a severe HRL show an increase by 13%. When comparing the unaided (no spatialization) and FM-only (spatialization only) conditions in both the moderate and severe HI subjects, 12 patients over 20 experienced an improvement or a preservation of their localization performance. The results with the HI subjects suffering from a profound HRL are even more satisfying, since 7 patients out of 10 experienced a decrease of their localization error with the spatialization applied in the FM-only mode. This enhancement goes from 12% to 75% in terms of performance. Finally, the localization errors observed in the aided and FM+M configurations does not markedly differ, indicating that the introduction of the FM-transmitted spatialized signal does not disturb the spatial hearing.

Localization and DOA

The analysis of the results has revealed some strong differences of performance between the sector locations. The localization error appeared to be the smallest in the frontal sector, according to several statistical variations between the central and intermediate/extreme sectors. This is in agreement with what is known about the spatial resolution of the AS, as reported in Appendix A.2.1. Many patients reported that it was difficult to make a choice between the sector 1, 2 and 3 (resp. -1, -2 and -3 for the left azimuths). It would be expected that the localization performance decreases as the source moves from the frontal azimuths to the lateral ones (see Appendix B.2). Yet, the outcomes show that the accuracy is better in the extreme sectors (L2/R2) than in the intermediate ones (L1/R1). This is most probably due to a bias in the protocol, because the participants could not give a perceived position beyond -4 and 4. In fact, several subjects reported that some stimuli came from some further directions, e.g. ±90°, and it this case their answer was ±4. With the headrest, people were barely able to see the ±4 numbers (that is, these numbers were almost out of the field of vision). If one would have added some ±5 possible responses, the patients should have been allowed to turn the head. On the other hand, the proportion of answers reporting a perception of the source from above or behind represented only 2.8% over the total of answers. Interestingly, 66% of this alternative answer arose from the profound HI patients, exhibiting a certain localization confusion, as discussed in the following.

Localization and group

The test reveals that NH subjects perform significantly better than HI subjects in an unaided configuration, even after SPL compensation. More precisely, the localization error of the NH group is multiplied by 3.4 in the HI-MOD group, by 3.7 in the HI-SVR group, and by 15 in the HI-PFD group. This indicates that a simple increase of the sound level is not sufficient to restore similar performance of localization in the HI patients, in agreement with what is reported in Appendices B.1.2 and B.2.1. When subjects were wearing their fitted HAs, this discrepancy dramatically diminishes, i.e. the localization error is multiplied by 1.1 (+12%) for moderate HI patients, and by 1.9 (+89%) for the severe HI patients. Only this last case found a

significant difference with the NH subjects. That result is similar in the FM+M configuration. One can conclude that the use of well-fitted HAs tends to reduce the gap of localization performance between the NH and HI subjects. This conclusion may be tempered by the fact that the NH subjects also experience a small degradation of their localization abilities with HA.

When the spatialization is rendered alone (FM-only configuration), the statistical analysis failed to show any significant effect between the NH and the HI patients suffering from moderate to severe HRL. Although it does not prove that the 3 groups perform the same, this somehow suggests that the significant degradation observed in the NH patients with the use of the spatialization feature would put the NH patients at a closer amount of performance compared to what was found in the unaided case.

Comparing the 3 HI groups, it seems that the moderate and severe HI patients provide some similar performance, except in the FM+M case, where a significant increase of localization error by 77% is found. This is in agreement with the information reported in Appendix B.2.1 that the impaired AS is capable to adapt and preserve some correct localization performance, especially for broadband stimuli. It should also be related to the fact that age has a well-established effect on the localization ability (see Chapter 4.3.1), and both groups predominantly include elderly. When the HRL exceeds a certain degree, one can see that the localization error explodes, as shown by the results obtained with the profound HI subjects. Such patients often reported that their better azimuth discrimination followed a left/center/right resolution only. With an average error of 1.5 to 2 sectors, one can recover this comment. This most probably explains the significant differences of localization performance that exist between the HI-SVR and HI-PFD groups, with an overall increase of the localization error by 187%. The reason for this discrepancy might not be only explained by the heterogeneity of PTAs in the HI-PFD group. The results also suggest that there exists a threshold of HRL beyond which the AS is not able to cope with such strong degraded localization information.

5.3.3 Preference-rating test

Overall results

Starting with the ideal spatialization quality, the proportion of subjects preferring the current rendering (i.e. “Diotic +” and “Diotic ++”) for the overall preference attribute decreases with the increasing HRL: 60% in the NH group, 50% in the HI-MOD group, 45% in the HI-SVR group, and 30% in the HI-PFD group. As expected, a majority of the NH listeners have no preference between the diotic and spatialized rendering concerning the intelligibility, and this is true for the 3 qualities of spatialization. On the contrary, the proportion of patients with a profound HRL that enjoy the functionality in terms of intelligibility is 45%, against 35% for the diotic condition. As reported by several subjects, this is related to the easier finding of the current speaker, which fastens the availability of lip reading. 55% of the NH listeners like the spatialization processing for the immersive feeling, whereas the average proportion

of “No preference” rating over the different HI groups reaches 37%. This suggests that the notion of immersion was unclear for a great deal of the HI patients, which was also observed by the examiner during the test. Over all attributes and HI groups, the preference for the novel functionality is 34%, against a 43% preference for the current rendering. This indicates that there is no noticeable plebiscite for the spatialization functionality.

Concerning the delayed spatialization, one can see that the total rejection (“Diotic ++”) in the NH listeners doubles and rises up to 35%, while it was only 17% with the optimal spatialization. This affects all the attributes, such as the overall preference for which the spatialization is enjoyed by 10% of the subjects, against 70% that does not like it. The same trend is noticed in the HI-SVR group, where the proportion of the preference toward the diotic rendering gets to 80% of the subject. This is most probably linked with the fact that the HI-SVR group is mainly constituted of elderly (see Appendix 5.1.1) that can more easily reject this spatial effect. In fact, such patients often indicated that they preferred a steady listening than a “moving” and imperfect rendering. This definite opinion is not observed in the moderate HI patients, where the indecisiveness is the primary answer (40% over the 5 attributes). Only the listeners suffering from a profound HRL still enjoy the processing, with an average 44% proportion of them preferring the new functionality over the 5 attributes. In particular, the preference in terms of intelligibility and pleasantness exceeds 50%. The first attribute is related to the easier lip reading. The second might be explained by the fact that 8 over 10 patients suffered from a congenital or infantile HRL. Such patients can be impressed by the spatial effect that they have never perceived before. No comparison with a natural spatial hearing can be done either.

Finally, the wrong spatialization condition brings about a clear refusal of the spatialization processing for the NH and HI-MOD groups, showing that the spatial effect is really perceived in both groups. This is also valid for patients with a profound HRL, for which the preference for the new functionalities goes down to 22%. In this category, the proportion of “no preference” is important, as well as the choice for a diotic rendering. Only the HI-SVR group ends with some curious results: the proportion of subjects that definitely rejects the functionality (mean of 20%) is less important that that observed with the delayed spatialization (30%). Again, this is probably due to the old age of the participants in this group. However, one should not conclude that the elderly are not sensitive to the binaural spatialization. Rather, this odd outcome must be linked to a higher risk of confusion (e.g. marking better the version 1, while the second is actually played and preferred). Also, with age, people often tend to be less consistent in their choice. Besides it must be mentioned that 6 subjects over 10 were older than 70, and the age of 2 of them exceeds 80.

Focus on the FM-experienced subjects

It is now suggested to split the HI patients into 2 subgroups: the ones with an experience in WMS (11 subjects), and the ones without (19 subjects). The difference in the preference-rating test between these 2 subgroups is impressive. First, a large majority of the FM-experienced participants (59%) enjoys the spatialization processing, which is not the case in the other

participants (31%). Moreover, the FM-experienced subgroup provides some very distinct results between the ideal spatialization and the 2 other qualities, which is not observed in the other subgroup. Indeed, the best quality yields a preference for the new functionality in 43% of the FM-experienced subjects (averaged over all attributes), against only 26% in the subgroup of non-experienced listeners. In particular 22% of the participants in the first subgroup really appreciate the spatialization, whereas they are only 7% in the other subgroup. Conversely, when the spatialization is wrong, the rejection is stronger for the FM-experienced patients (60% rejection, with 30% "Diotic ++") than for the non-experienced subjects, who present a rejection of 50%, in which only 13% reported the "Diotic ++" evaluation. The same observation can be made when looking at the "overall preference" attribute only. For the FM-experienced patients, the balance between the clear preference with the best quality and clear rejection with the worst quality is 41% vs 32%. In the non-experienced subjects, the balance is 11% vs 17%. Note that the notion of immersion is still unclear in both subgroups.

2 main conclusions must be drawn. First, the patients experienced with the technology of the WMS strongly enjoy the new functionality over the present one. However, this is only true if the spatialization is ideal. That is, the current system would not be appreciated with its current performance. Second, the comparison between the spatialized and diotic renderings should preferably be conducted with people that are using, or have used, the current WMS. This makes sense, because the other subjects never encountered the issues of diotic speech, i.e. they never suffered from the absence of spatial hearing with such types of solutions.

If one focuses on the 2 attributes depicted on Figure 5.18 in the static ideal scenario, the results are flattering. Concerning the naturalness, a total of 64% of subjects appears to prefer the new functionality, and the proportion of "Spatialization ++" reaches 55%. In this subgroup, 8 over 11 HRL came from a congenital or an infantile disease. A great deal of those patients never experienced a normal spatial hearing. For them, the introduction of an ideal spatialization brings back what would be a natural perception of the sound. When it comes to the overall preference, the same proportions can be observed. Additionally, the amount of patients rejecting the spatialization falls to 18%. This sounds like a very promising result for WMS including a spatialization functionality.

5.4 Conclusion

The core contribution of the thesis has been reported in this chapter. The evaluation of a binaural spatialization processing on aided HI subjects is a hitherto unseen study, which is prominent for the future development of HAs, and particularly WMS. Indeed, the interest in binaural algorithm and spatial rendering via HAs is constantly growing, but it is unclear how aided users react to these renderings. At the beginning of the research it was unsure whether the desired functionality could be useful and appreciated by HI patients. The reported results show that it makes sense to keep up investigating this topic for further researches and developments.

Additionally to the results and discussions about the clinical trial, this chapter is also based on an extensive review (Appendix E) of the literature concerning the evaluation of various auditory features on both NH, aided and unaided HI subjects. This thorough investigation in the state-of-the-art has been compulsory to build the test protocol and submit it to an ethics committee. Also, it has allowed to rapidly identify the risks for the patients and possible biases of the results. Moreover, the collaboration between the author, the involved audiologist, and Phonak has been quite efficient. For instance, 50% of the targeted HI patients have accepted to take part to this clinical trial. Thanks to all these elements, the ethical validation of the protocol has been obtained after a few months only. The literature has appeared interesting and useful for the decisions related to the following items:

- The sample size. The average number of 20 participants reported in the literature has shown that there is the necessity to have a sufficiently large number of subjects, to end up with statistically reliable outcomes. On the other hand, the available subjects that can take part to such a study is not infinite. The requirements to work with HI patients with different degrees of HRLs, exhibiting a symmetric HRL, and being equipped with bilateral BTE Phonak HAs is highly selective. This had to be taken into account when fixing a certain targeted number of participants,

- The age of the tested patients. In particular, the fact that only 1 study reported the presence of children in the panel has motivated the resort to adult subjects only, even though children are actually the main end-users of this functionality. Ethical considerations has also supported the absence of children in the clinical trial. To partly circumvent that issue, it has been decided to include some young adults, preferably experienced with WMS,

- The management of HAs and their related signal processing features. The review of the previous studies has evidenced that there is no conventional rules, and that the decision to activate or deactivate certain algorithm only depends on the goal of the research,

- The stimuli. With the availability of meaningful or unpredictable speech database, one had to think on the most appropriated material. Moreover, the choice of the speech-shaped noise has came from what is reported in the literature,

- The procedure of the different tests. This primarily concerns the use of the SRS for the assessment of the speech intelligibility. The procedure of some adaptive SNRs has been adopted for the clinical trial. Concerning the localization, the resort to a head rest and the limitation of the learning effect have been supported by the literature as well. Finally, the choice of certain sound attributes, and the way how the participants should report their preference have been inspired by several previous studies.

The major results of the clinical trial are summarized here:

1. The spatialization, as developed in this thesis, does improve the intelligibility performance of some categories of subjects, compared to the ones reached with the current

WMS. That is, there is no negative consequence of applying HRTFs with limited magnitude to HI subjects in terms of speech understanding, although this intrinsically introduces an ipsilateral ear and a contralateral ear, and reduces the binaural summation,

2. The use of generic, amplitude-limited, bandwidth-limited, and approximated HRTFs would have no detrimental effect on the localization ability of HI listeners in the FHP, contrary to NH listeners,
3. It might exist a certain HRL threshold above which the localization of sound sources is suddenly arduous,
4. HI subjects are subjectively sensitive to binaural spatialization,
5. Subjects with an experience in the use of WMS significantly appreciate the spatialization functionality, as long as it is fast and right,
6. The preference of features related to WMS should be preferably assessed on subjects experienced in this solution,
7. The spatialization functionality brings the perception of a more natural sound in HI subjects suffering from a congenital or an infantile disease-induced HRL. They also find it more pleasant than a diotic rendering,
8. Lip reading is made easier with the spatialization functionality, which is found quite useful by HI subjects, especially the persons suffering from a profound HRL,
9. With its current performance, the developed system will not satisfy the patients.

Every clinical trial presents some advantages but also some limitations. Table 5.11 draws up a list of what is thought to be the major strong and weak points of this study.

As a conclusion, one can state that the reported clinical trial may serve as a reference for further developments in the HA industry. It also provides some insights in the way how the processing can be improved, in order to fulfil the subjects' requirements.

Strong points	Weak points
The sample size is large and well-balanced between groups. This justifies the generalization of the outcomes to a larger population.	The results are valid for HI subjects presenting a relative symmetrical HRL. They should not be generalized to any kind of audiogram patterns.
The study includes various origins of HRL (congenital, disease-induced, presbycusis...) and several generations of participants.	The masker played in the intelligibility test (speech-shaped noise) does not correspond to any real-world noise. The rendering of the noise via the DAI, rather than the HA microphone, is also unrealistic.
The panel is composed of a subgroup of 11 subjects experienced with the diotic rendering provided by the current WMS.	The resort to a headrest does not take into consideration the localization as done in the daily life.
The primary signal processing features of the HAs, especially the amplitude and frequency compressions, have been kept activated, although they distort the added spatial cues. Moreover, one has taken into consideration the dynamics of the fitted HA to ensure that all the desired SNRs were correctly rendered.	The forced choice between several spatial locations, as well as the absence of possible answer beyond ±65° introduce a certain bias to the localization test results.
The clinical trial reports some outcomes that cover various evaluations (intelligibility, localization...) and effects of numerous factors (SNR, rendering, DOA, modes) on the subjects' performance.	Despite the systematic introduction of a training period, a possible long-term learning effect during the test may have brought about some consequences on the outcomes.
The collection of the patients' feedbacks, as well as the use of both some objective and subjective evaluations, allow to get a wide overview of the advantages and limitations of the current spatialization functionality.	The duration of the test was long (90 minutes) and could possibly result in a certain fatigue and a decrease of the concentration, even though some mandatory breaks were demanded.

Table 5.11 – List of the strong and weak points of the clinical trial.

Conclusion

This thesis has proposed a solution to restore spatial hearing in WMS, which are known to significantly increase the speech understanding performance of the equipped HI listeners. The new spatialization functionality has been implemented on an embedded prototype, and evaluated via an extensive clinical trial. Its feasibility and benefits have been evidenced throughout the thesis. This novelty falls within the scope of a trend aiming at taking advantage of binaural hearing towards the enhancement of the current wireless HAs. In particular, it is interesting for the processing performed in the WMS and beamforming algorithm. The research about this functionality development has shown that numerous well-established signal processing methods can be adapted to the technical constraints demanded by HAs. Nevertheless, they must be shrewdly modified to take into account the characteristics of the disabled AS.

Providing new capabilities to HAs may increase the adoption of these solutions by subjects suffering from HRL. Also, it might support the spread of binaural HAs and bilateral fittings, against monaural renderings and unilateral fittings, which do not benefit from the binaural hearing property of the AS. The detailed results and conclusions, as well as certain directions for future research, are presented hereafter.

Results and original contribution

Algorithms for binaural localization

Despite some strong technical limitations from the hardware and software (low memory storage and computational power, absence of streaming between both devices, limited buffering at the input, necessity for a real-time processing...), the thesis has demonstrated that a binaural localization algorithm can be integrated in HAs. This is enabled by an optimized combination of signal processing methods of low complexity, which enable a fast-acting algorithm. The availability of several cues from different domains (acoustics and electromagnetics) does not yield a very precise angular resolution, but strongly contributes to the prevention of strong localization errors (e.g. localizing a source on the right while it is on the left). These kinds of mistakes would be indeed dramatically prejudicial for the future acceptance of the new functionality by HI subjects.

Binaural spatialization for hearing aids and hearing-impaired persons

Well-established techniques of binaural spatialization has shown to be appropriate for HA purpose, provided that they are adapted to those devices (limited computational power and memory, restricted bandwidth, presence of signal processing features that deteriorate the rendering of spatial cues...). The design of low-order filters and the implementation of a simple interpolation scheme guarantee the real-time processing of the algorithm. The application of spatialization methods to HI subjects is at its early stages. The introduction of the concept of HRTF dynamic limitation is a step toward this direction. Although it has not been specifically tested on HI listeners, it appears to be consistent for NH listeners, who are quite a bit more sensitive to the artificial spatial rendering.

Prototype development and evaluation

The optimization of the BLA has shown to be complex, due to the numerous tunable parameters, and the strong dependency of its performance on the setup. Despite the computation of some advanced optimization schemes, only light improvements in terms of accuracy and reaction time have been observed after this operation. The porting on an embedded device has been successful thanks to a close control of the fixed-point variable resolution and automatic C-code generation. The final assessment reveals that the BLA is quite accurate but a bit slow in real acoustic environments. This is the major weak point of the algorithm, that can be improved in some further researches.

Clinical trial for evaluating binaural spatialization on disabled subjects

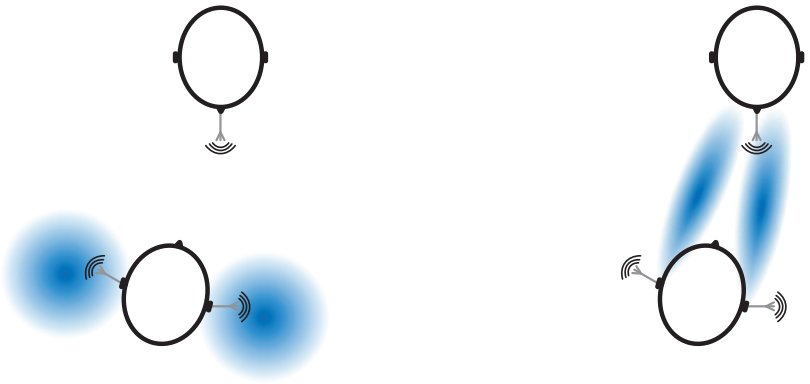
The extensive clinical trial, incorporating a large panel of patients, has disclosed some interesting observations and facts. First, the novel spatialization functionality improves or preserves the speech intelligibility performance, even though it removes a great part of the binaural summation advantage from current WMS. Second, the use of generic-based HRTFs and approximated spatial filters, of which the magnitude is limited, has a marginal effect on the localization abilities of HI subjects in the FHP, whereas it is detrimental for NH listeners, as could be expected. Finally, the disabled tested subjects that are experienced with the use of WMS clearly enjoy the new functionality, as long as it is accurate and fast. The resolution of 5 spatial sectors does not have any effect on the perception of moving sound source. These conclusions validate the initially stated hypothesis that including spatialization in WMS can be beneficial for HI persons.

Future research

The results from the clinical trial strongly support the continuation of the resarch conducted in this thesis. Various topics can be concerned by some further improvements and investigations.

Binaural localization algorithm

Considering the final performance of the BLA, one can think of several possible future researches. Substantial improvements can come from new hardware and software capabilities. Since the available processing power is always growing in electronic devices, more computations will be enabled in future HAs. One of the most promising novelty is the availability of 3 audio inputs in the HAs. Previously, there were only 2 inputs, so that one of the 2 HA microphone had to be deactivated when using WMS (i.e. one of the 2 inputs was devoted to the wirelessly-transmitted audio signal). The consequence was that no beamforming could be processed in the HAs (Figure 5.19A). In recent and future hearing devices, the presence of 3 inputs (2 audio from the HAs + 1 audio from the WMS) makes possible the restoration of directivity, like in the standard use of HAs (Figure 5.19B). Therefore, the resemblance between the audio and radio-transmitted signals is quite higher, which should allow for a more efficient and precise binaural localization.



(a) (b)

Figure 5.19 – Comparison between the previous omnidirectional microphone mode (A), and the directivity that is now available in the HAs using WMS (B).

Here are some other ideas for enhancements:

- The possibility to buffer signals on a longer duration. This would bring more freedom to compute the IACS and integrate some methods from the IACC-based algorithms, such as the generalized cross-correlation. It would allow to develop a process that has been discarded in the current BLA: the search of similarities between the signals from the body-worn microphone and the one from the HA microphones. The objective is that only the information from the speaker are extracted from the left and right degraded signals, so as to achieve the localization processing on reliable time-frequency segments only. Additionally, this possibility would enable the estimation of the distance between the speaker and listener, which was reported as a subsidiary functionality in Chap 1.4.3,

- In the near future, the exchange of full audio frames between a pair of HAs will be made easier and less power-consuming. This would facilitate the comparison between the left and right signals,

- More and more BHAs integrate environment monitoring algorithms to automatically adapt the current program of the HAs. The information from such algorithms could be transmitted to the BLA, in order to complement or replace the current intermodal coherence estimation. The same concept applied to assess the RF surroundings is another interesting idea,

- The use of some advanced signal processing methods. The techniques based on sparsity or the resort to neural networks, which could learn the characteristics of numbers of acoustic environments and the related quality of the spatial cues, could be considered. This goes with the trend to develop intelligent hearing devices [63, 86, 89],

- The integration of advanced tracking methods, such as the particular filtering that has shown to give great performance in another localization algorithm developed at the LTS [148].

The evaluation of the current BLA in various noisy conditions could be the framework of the development of such suggestions. Right now, it is unsure whether it is relevant to increase the spatial resolution of the current system. On the one hand, the current 5-spatial sector resolution appears to be sufficient for HI subjects, after the results of the clinical trial. On the other hand, a finer algorithm could yield a faster localization and a lower interpolation time.

Binaural spatialization algorithm

Many of the possible improvements concerning the BSA are primarily guided by the enhancement of the BLA. The current BSA adds some latency when a speaker's motion has to be rendered through the procedure of HRIR interpolation. If the BLA could achieve a higher resolution, a larger HRTF database could be stored and the recreation of movements could be performed faster. This is to address the necessity for real-time computations. When it comes to the offline processing, an advanced and optimized introduction of magnitude range limitation could be noteworthy. The idea would be to combine a mixture of dynamic compression and limitation with different ratios and thresholds to simplify the shape of the HRTF, while preserving at most the perception of a suitable spatial hearing. This might allow a subsequent filter order reduction.

Technical aspect

The primary following step is to integrate the developed algorithms on a pair of HAs, which will replace the current prototype that requires wires and centralizes the computations. The related questions are:

- How to efficiently share the processing between both HAs?
- How to manage the slight clock differences in the 2 devices, as well as the internal jitters?
- How to adapt the code so that it works on a DSP rather than on a processor-based microcontrollers (e.g. the current Atmel ARM9)?

Subjective assessment and clinical study

This thesis deals with the thematic of the redering of spatial audio with HAs. It is just the beginning of some further explorations within this research topic. The conducted clinical trial gives birth to several other noteworthy studies related to:

- The perception of adult-HRTF based spatial filters on HI children (i.e. the major end-users of the new functionality). Their localization performance might be potentially reduced due to the difference of head size,
- The concept of externalization in HI listeners. It is not impossible that certain categories of HI subjects do not perceive a difference between a HRTF-based spatialization and a simple lateralization (i.e. the application of frequency-independent ITD and ILD). This would mean that either the current BSA does not bring a sufficiently faithful spatial rendering, or that HI listeners are not sensitive to externalization, presumably because of their limited access to monaural cues. Whatever the rationales, such an observation could dramatically simplify and fasten the processing done in the BSA,
- The benefit from the current spatialization on patients presenting an asymmetrical HRL, provided that they have a bilateral fitting. Such people exhibit some very degraded localization performance, and it would be interesting to know whether the BSA could enhance their abilities,
- The long-term effect of the spatializaton functionality, in order to check if the daily learning performed by the CAS can outperform the usual aided localization performance of HI persons,
- The effect of the BSA on speech understanding with some spatially separated noise-types and speech. This is equivalent to determine the SRM provided by the new functionality, compared to the SNR of the everyday-life of the same HI subjects. An improvement of the speech intelligibility is expected,
- The localization ability of HI patients in noisy and/or reverberant environments, with and without the new functionality. Again, an enhancement of the localization performance is expected.

The author suggests to conduct both technological and clinical researches simultaneously, since they are intrinsically related to each other.

A Appendix: Normal hearing

Here are described the human AS properties, in terms of anatomy, sensitivity, frequency and temporal resolutions. Then, the notion of binaural hearing is investigated, and the subsequent effects on sound localization and speech intelligibility are reported.

A.1 The auditory system

A.1.1 Peripheral and central hearing structures

The AS is composed of 2 major structures: the *peripheral* and the *central auditory systems* (resp. PAS and CAS). The PAS is constituted of the outer ear, the middle ear, the inner ear, and the auditory nerve, which ensure the connection between the peripheral structures and the auditory cortex (CAS). Figure A.1 depicts the PAS in detail. The main role of the ear is to convert an acoustic signal in an electrochemical stimulation, via several consecutive transductions: acoustic/mechanic, mechanic/hydrodynamic, hydrodynamic/electrochemical.

The pinna and the external auditory canal are the primary components of the outer ear. The pinna, head and torso, alter the sound amplitude and introduce some delay in a frequency-dependent manner. As the pinna structures are small, they only affect the HFs (i.e. short wavelengths), while the lower frequencies are affected by the head and torso. Apart from its protective role, the outer ear has a function of amplification (sound-gathering effect), and plays an important role in sound localization, as discussed in Appendix A.2. Given the size of the concha, located at the input of the ear canal, and the length of the ear canal, the outer ear brings a gain on the acoustic pressure between 1.5 to 7 kHz. This results from the combined resonances of the auditory canal (around 2 kHz) and the concha (around 5 kHz), as shown on Figure A.2.

The middle ear is composed of the eardrum, the 3 ossicles (malleus, incus and stapes) and the eustachian tube. The latter is connected to the nasopharynx, and ensures the pressure equalization between the environment and the middle ear cavity. The role of the middle ear is

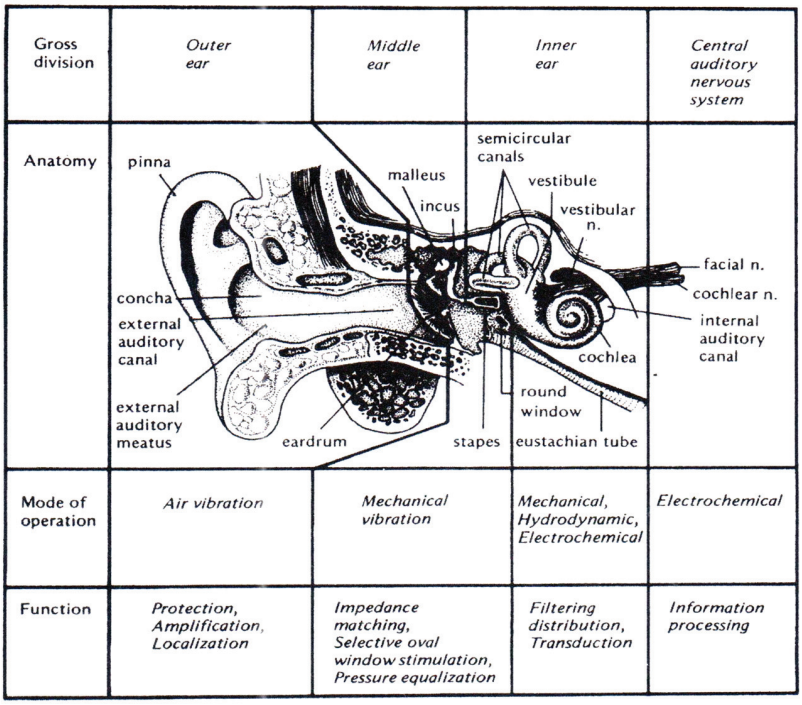


Figure A.1 – The PAS, made of the outer, middle and inner ear, and the auditory nerve. From [250, page 68].

to help overcome the impedance mismatch between the outer ear and and inner ear. Indeed, a direct connection between the 2 would be highly inefficient, since the latter is composed of fluids and tissues denser and stiffer than the air (i.e. of high impedance). If this impedance adaptation did not exist, 99.9% of the airbone energy would be reflected back and only 0.1 % could be transmitted to the inner ear [78, Chap. 3]. There are 2 main pathways that bring sound to the middle ear. The first one is the conversion of the air vibration in a mechanical stimulation via the tympanic membrane. The other is the bone conduction resulting from the skull vibrations. Those vibrations are partly transmitted directly to the inner ear via the cartilaginous path [253]. The other part is radiated in the ear canal.

When the air conduction is blocked (e.g. by means of a HA), the bone-conducted sound cannot escape outside the ear canal [171, Chap. 9]. It results in an artificial amplification of the LF, which can be disturbing and uncomfortable [78, Chap. 3]. This effect is called the *occlusion effect*. It is not perceived when the air transmission is free because the ear canal acts as a high-pass filter. Finally, the *acoustic reflex* must be mentioned, as a mechanism to protect the ear from excessive loud sounds. It consists in a contraction of the middle ear muscles that provides an intensity reduction from 10 to 30 dB. The major limitation of this reflex is its low reaction time: around 150 ms for a 80 dB stimulation, down to 10 ms for higher levels. The prevalence of HRLs due to too loud sounds shows that the ear does not have an adequate and efficient protective mechanism [250, Chap. 6].

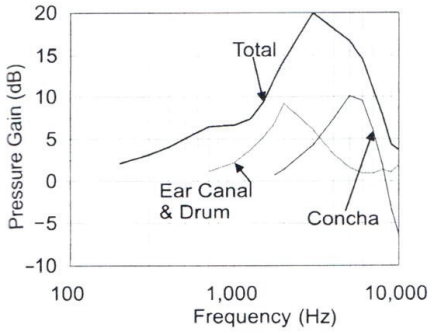


Figure A.2 – Transfer function of the outer ear, resulting from the combined resonances of the ear canal and concha. From [250, page 73].

The components of the inner ear are the oval window, the vestibule and the vestibular apparatus. It is located in the temporal bone of the skull. The vestibular apparatus is responsible of the sense of balance, spatial orientation and acceleration detection, and is not discussed in this thesis. The oval window is the interface between the middle ear and inner ear. It is excited by the motion of the stapes and stimulates the fluids located in the *cochlea*. The latter consists in a tube of decreasing diameter that is coiled increasingly sharply on itself. It is constituted of 3 ducts: the scala vestibuli, the scala tympani and the scala media. The *basilar membrane* acts as a separation between the scala media and both the scala vestibuli and scala tympani, as can be seen on Figure A.3A. The scala media contains the endolymph, with an electrical potential of 80 mV, while the scala vestibuli and scala tympani accommodate the perilymph at a null potential. This electrochemical difference is the basis of the generation of neural activity inside the cochlea.

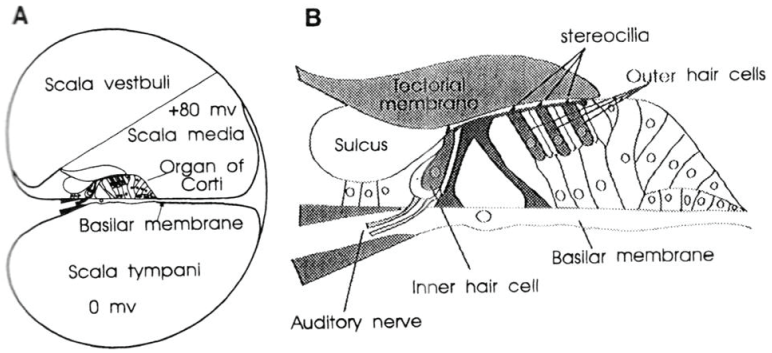


Figure A.3 – Section of the cochlea (A) and section of the organ of corti (B). From [170, page 45].

A section of the *organ of Corti*, which is situated between the basilar and the tectorial membrane, is shown on Figure A.3B. The outer side contains 3 or 4 rows of *outer hair cells* and the inner side accommodates the *inner hair cells*. *Stereocilia* are located at the upper surface of both types of cells. The cochlea is made of about 3500 inner hair cells with 40 stereocilia each,

while it counts 12000 outer hair cells with 150 stereocilia each. All these cells do not regenerate and their number usually decreases with age (see part B.1). The movements of the endolymph set the hair cells in motion. The behavior of the outer and inner hair cells is different. The outer hair cells expand and contract, making a variation of their size. Thus, they provide an amplification of the inner hair cell movement that intensifies their response. The bending of the stereocilia brings about neural discharges in the auditory nerve. One particularity of the hair cells is that they are highly selective in frequency. Moreover, the variations of width and stiffness of the cochlea cause maximum vibration at different stages along the basilar membrane. Therefore, the cochlea can be viewed as a bank of bandpass filters, where the HFs are treated at the base and the LFs at the apex, as depicted on Figure A.4. Note that there is no capillary inside the basilar membrane, so that the sound associated with blood circulation is attenuated.

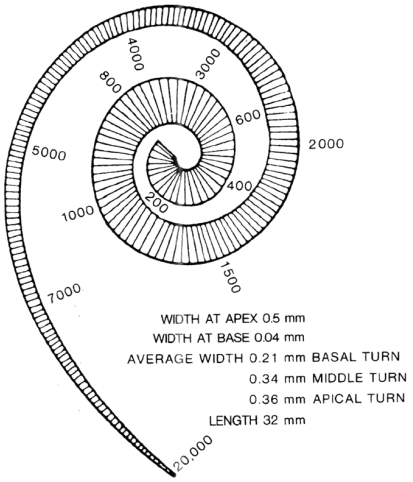


Figure A.4 – Frequency selectivity along the cochlea. From [241, page 12].

The stereocilia convert the movement of the inner hair cells in electrical excitations in the neurons to which they are connected. These neurons constitute the auditory nerve. A neuron is characterized by its spontaneous activity, which is its response to the absence of stimulation. The physiological response translates into a discharge and creates a neural spike. The discharge rates is the number of times a neuron generates a spike in a given time. It has been shown [250, Chap. 9] that the sound intensity is coded in the auditory nerve through the discharge rate (i.e. louder sounds yield higher rates). The neurons with a high threshold of spontaneous activity encode the low *sound pressure levels* (SPLs) (i.e. from 50 to 250 spikes/s between 0 to 20 dB SPL), while the ones with a lower threshold of spontaneous activity encode higher SPLs (i.e. from 50 to 250 spikes/s between 20 and 80 dB SPL). This enables the AS to cover a great range of SPLs. The different neurons do not discharge at a similar stimulation rate, which allows to encode the frequency. The level of the tone required for the neuron to go over its spontaneous activity as a function of the frequency is called the tuning curve. The tuning curve of the neurons matches the frequency selectivity in the basilar membrane (see

e.g. [78, page 185]).

A.1.2 Hearing thresholds and scales

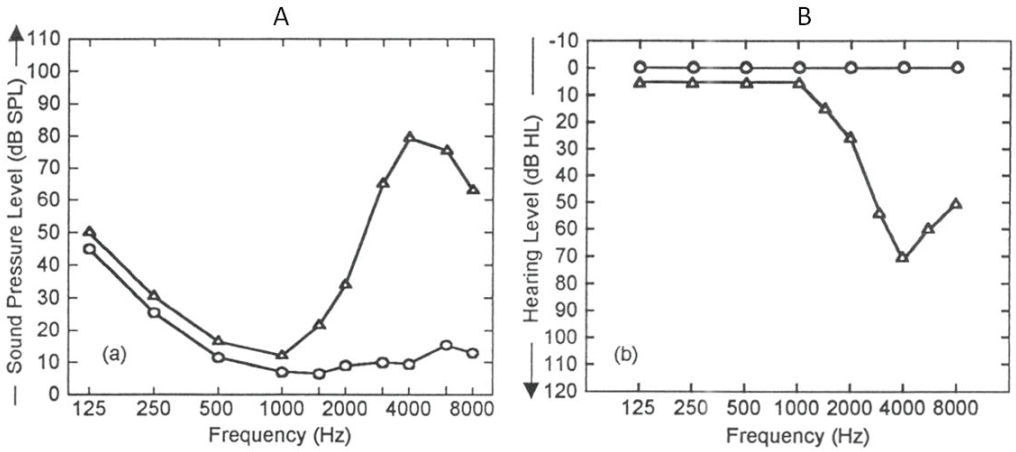


Figure A.5 – The thresholds of hearing for a NH subject (circles) and a HI subject (triangles) (A). The corresponding audiograms are shown on the right panel (B). [78, page 338].

Theoretically, human beings can hear sounds at frequencies between 20 Hz and 20 kHz. The upper boundary actually decreases rapidly with age, down to around 16 kHz. The sensitivity of the ear depends on the frequency. The *threshold of hearing* (or absolute threshold) is defined "*for a given listener as the minimum SPL of a specified sound that is capable of evoking an auditory sensation*" [150]. In *tonal audiometry*, these stimuli are sinusoids at different frequencies. The determined hearing thresholds can be plotted as a function of frequency, as shown on Figure A.5A. It represents the sensitivity of the ear and can be seen as the *transfer function* (TF) of the AS to tonal sounds. The circle markers correspond to a typical MH subject, while the triangle markers show an example of a HI subject, which is discussed in part B.1. In order to get the so-called *audiogram* of a subject (Figure A.5B), a normalization to the reference of normal hearing thresholds is performed. Hence, the hearing levels are expressed in dB HL instead in dB SPL, e.g. 0 dB HL at 1000 Hz corresponds to 7 dB SPL. Not that the dB SPL is also referenced by the threshold of human hearing (20 μPa), but it does not take into account the frequency-dependent audibility in detail.

The *loudness* of a sound is "*an attribute of auditory sensation in terms of which sounds may be ordered on a scale from soft to loud (IEC 801-23-05)*" [150]. It is expressed in *phon*, which is "*numerically equal to the sound pressure level of a 1 kHz tone which is judged to be equally loud*" [150], i.e. x dB SPL = x phons at 1 kHz. When the intensity increases, the subjective perception of sound level increases as well, but there is not a one-to-one correspondence between the 2 [172, Chap. 4]. Loudness allows to express this perception. *Equal-loudness contours*, or phon-curves, are the results of loudness measurements on a large panel of subjects (see e.g. [172, page 135]). They are used to define units that consider the human perception of intensity.

Such units are the dB(A), dB(B) and dB(C). One noticeable property of the loudness is that its contours become flatter at high SPLs. For instance, the loudness dynamics equals about 70 dB between 20 Hz and 20 kHz for lower sound intensities, while it falls to about 30 dB at high sound intensities. The A-weighting is referenced to the equal-loudness contour at 40 phons and the B-weighting is for sound at around 70 phons. Louder sounds require the C-weighting. In practice, only the dB(A) is used in psychoacoustics, whatever the SPL. Another scale for loudness is the sone scale, but it is not employed in this thesis.

Loudness is also used to assess the thresholds that can lead to hearing damages. While the *uncomfortable thresholds* can reach 120 to 140 dB in pure tone audiometry, quite a bit lower levels are already dangerous for the ear. In addition to the loudness, another important factor is the duration of exposure. In Australia, the National Acoustic Laboratories [129] recommend not to exceed 1 minute a day at 110 dB(A). Then, the sound intensity decreases by 3 dB each time the duration is multiplied by 2 (e.g. max. 2 minutes at 107 dB(A), max. 4 minutes at 104 dB(A), and so on). Note that a typical conversation is around 60-70 dB(A), while the music in a discotheque is rendered at levels that can reach 105 dB(A). Finally, it must be mentioned that the loudness decreases over time for a sound delivered at the same intensity [250, Chap. 13]. This phenomenon is called *loudness adaptation*.

A.1.3 Frequency processing

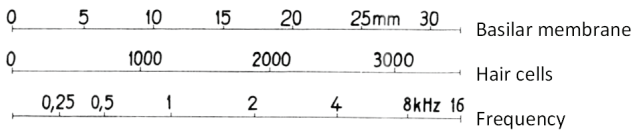


Figure A.6 – The uncoiled cochlea from the apex to the base and the corresponding numbers of inner hair cells, as a function of the frequency. Adapted from [254, page 89].

As previously said, the basilar membrane of the cochlea includes the hair cells that are sensitive to different frequencies, depending on their location. These cells are distributed along the organ of Corti, as depicted on Figure A.6. This picture is a diagram of the uncoiled cochlea from the apex to the base, and shows the relation of these cells with the frequency. It can be noticed that about 1000 inner hair cells are devoted to a bandwidth of 1 kHz, while the next group of 1000 hair cells are in charge of a 2 kHz-bandwidth. The last 1500 inner hair cells cover the frequency range from 3 to 16 kHz. That is the reason why the AS presents a finer frequency resolution in the LFs. The resolution decreases with the frequency, and the frequency processing of the ear follows a log-like scale rather than a linear one.

It has been previously reported that the basilar membrane behaves like a bank of band-pass filters. Their overlapping passbands are called the *auditory filters*. The shape of such filters can be determined by using some masking techniques. *Masking* is defined in [150] as "*the amount by which the threshold of hearing for one sound is raised by the presence of another sound*". It is

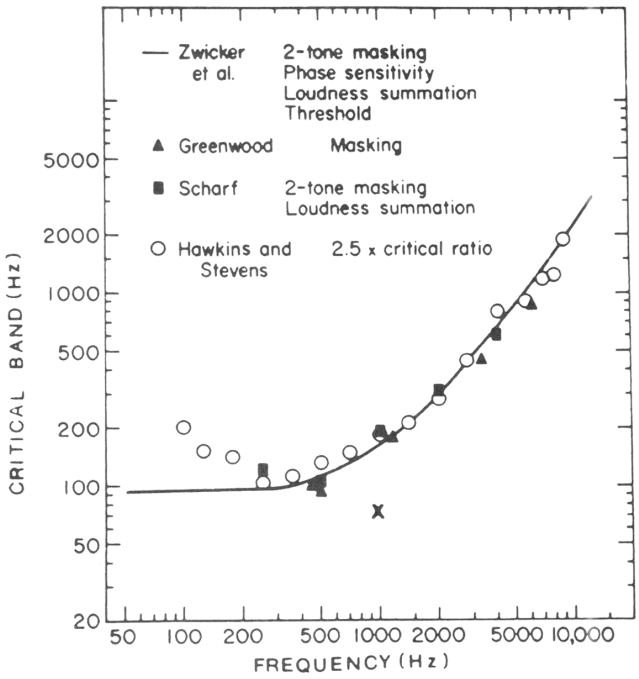


Figure A.7 – The auditory filter bandwidth that enlarges with the increasing frequency. From [78, page 315].

expressed in dB. When listening to a pure-tone signal in a background of noise, the basilar membrane creates an auditory filter of which the center frequency is located close to the main frequency content of the signal. This suppresses a great part of the noise that does not cover the corresponding bandwidth. An important property of the AS has come across with this method. When increasing the bandwidth of the noise, the portion of the noise passing through the auditory filter increases up to a certain bandwidth, above which the amount of noise going through the filter does not change any longer. This bandwidth is called the *critical bandwidth*. It increases with the frequency, as it can be seen on Figure A.7. Making the assumption that the bandwidths of the auditory filters have a rectangle shape, one speaks of a *critical band* instead [172, Chap. 3].

Masking curves show the thresholds of hearing for increasing levels of white noise (see e.g. [254, page 54]). 2 main properties are highlighted in those curves. First, the masking effect is small in LFs. It then steps up by 10 each time the frequency is multiplied by 10 dB. For instance, at 40 dB of noise, the masking equals 20 dB in the 50-1000 Hz bandwidth, while it reaches 30 dB at 10 kHz. Second, the masked thresholds tend to flatten when the noise SPL augments. When understanding speech in noise, it has been found that the AS is able to select the bandwidth with the best *signal-to-noise ratio* (SNR). This is discussed in part A.2.2, where the notion of speech intelligibility is introduced.

A.1.4 Temporal processing

The *temporal resolution* of the AS is defined as "*the ability of a [...] person to distinguish the temporal structure of a signal.*" It "*may be evaluated in terms of the fastest timescale on which a signal element such as a gap can be detected*" [150]. The average gap detection time is around 2 ms. This was determined by various studies reviewed in [78, Chap. 9]. It is independent of the SPL and of bandwidth of the stimuli, but it depends on their duration. In detail, it goes from 1 ms to 50 ms for stimuli lasting 0.5 s to 1 ms. The temporal resolution is related to the *envelope* of the stimuli. It has been shown that the resolution is independent of the modulation frequency below 16 Hz. Then, it reduces quickly for upper modulation rates [172, Chap. 5]. Eventually, note that the temporal resolution must not be confounded with the *temporal integration*, which is "*the ability of the AS to add up information over time, to enhance the detection or discrimination of stimuli*" [172, Chap. 5]. The average integration time of the AS is about 80 ms.

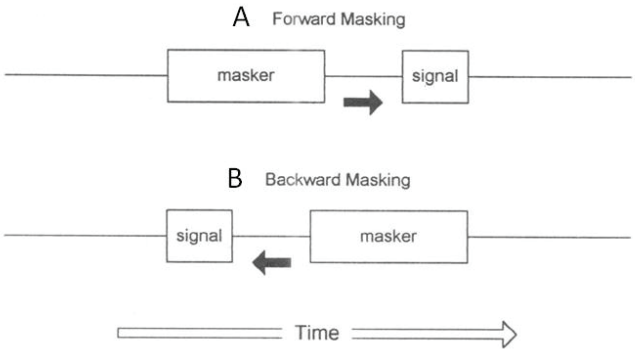


Figure A.8 – Temporal masking: forward masking (A) and backward masking (B). Adapted from [78, page 326].

Masking occurs in the temporal domain as well. It is called the *temporal masking* or non-simultaneous masking, and results from the temporal resolution of the ear and its "latency". In this case, there is no temporal overlap between the masker and signal. *Forward masking* takes place when the masker is presented first, stopping a brief time before the signal (Figure A.8A). The shorter the delay, the higher the masking. Indeed, the masking increases by 30 dB when the gap duration is decreased from 25 ms to 1 ms, when only one ear is stimulated (monotic rendering) [78, Chap. 10]. This difference reduces to 5 dB when the rendering is dichotic (the signal on a side, the masker on the other). This is a first insight into the advantage of binaural hearing, which is detailed in part A.2. *Backward masking* concerns the case when the masker appears just after the end of the signal (Figure A.8B). When the delay varies from 1 to 50 ms, the masking is diminished by 20 dB for monotic rendering, whereas the reduction is 2 dB in dichotic rendering, as the masking is already low. The increment of the masking when the masker SPL increases is small, around +3 dB when the masker level is stepped up by 10 dB. In parallel, the frequency selectivity is shown to augment in backward masking, i.e. the bandwidth of the auditory filters decreases, resulting in an efficient suppression mechanism

[170, Chap. 5].

The physiological causes of the temporal masking are not well understood. Several hypotheses exist and are related to the cochlear and central processing. The forward masking might be due to a persistence of the masker representation in the auditory nerve. Concerning the backward masking, a kind of overriding phenomenon might occur before the signal has been fully processed [78, Chap. 10].

A.2 Binaural hearing

A.2.1 Sound localization

When it comes to object localization, vision is known as the most reliable sense used by the brain. Indeed, the performance of localization based on vision is twice better than the one based on acoustics [22, Chap. 2.1]. However, vision does not allow to infer the position of objects when they are out of sight. In these situations, sound localization becomes the only resort to locate objects in space. Note that for HI subjects, the combination of visual and acoustic cues facilitates the lip reading, a powerful technique to help speech understanding. In this thesis, only the localization in the horizontal plane is covered because the newly developed algorithms presented by the author only concern this type of localization.

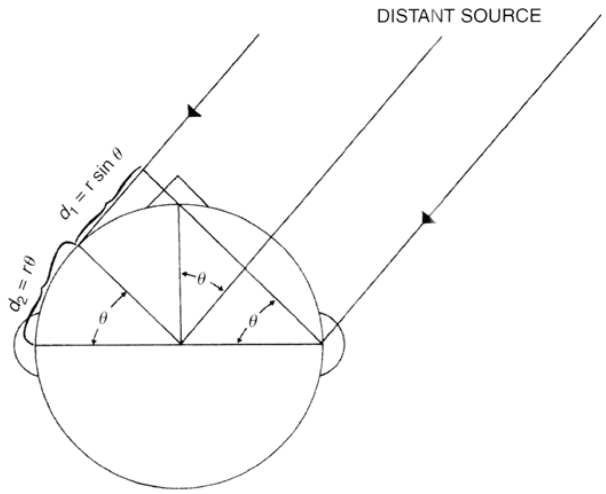


Figure A.9 – A situation where the acoustic waves of a distant source arrive to the ear of a listener. The angle θ is taken positive when the distant source is on the left. From [241, page 37].

Binaural and monaural hearing provide the ability to locate sounds in space. This is possible through the use of *binaural cues*, *monaural cues* and *dynamic cues*. Let consider the situation depicted on Figure A.9, where sound waves coming from a given direction reach the head of a subject. Far field conditions are hypothesized, i.e. wavefronts correspond to plane waves. The

wavefronts arrive at the *ipsilateral ear* first and attain the *contralateral ear* after a certain delay. This delay is called the *interaural time difference* (ITD). It appears when the wavelength of the sound is larger than the distance of the curved path between the ipsilateral and contralateral ears, that is for frequencies below around 1.5 kHz [78, Chap. 13]. In higher frequencies, a phase ambiguity occurs and prevents the AS from resorting adequately to the ITD. Indeed, it is impossible to determine which cycle of the *temporal fine structure* (TFS) at the contralateral ear corresponds to a given cycle at the ipsilateral ear.

Considering the head as a sphere, the ITD δ can be computed as a function of the incidence angle θ with the Woodworth's formula [249, page 396]:

$$\delta(\theta) = \frac{d_1 + d_2}{c} = \frac{r}{c}(\theta + \sin\theta), \tag{A.1}$$

where c is the sound speed in the air and r is the radius of the head. Ignoring the path around the head, equation A.1 is simplified in the *sine law* [238]:

$$\delta(\theta) = \frac{a}{c}\sin\theta, \tag{A.2}$$

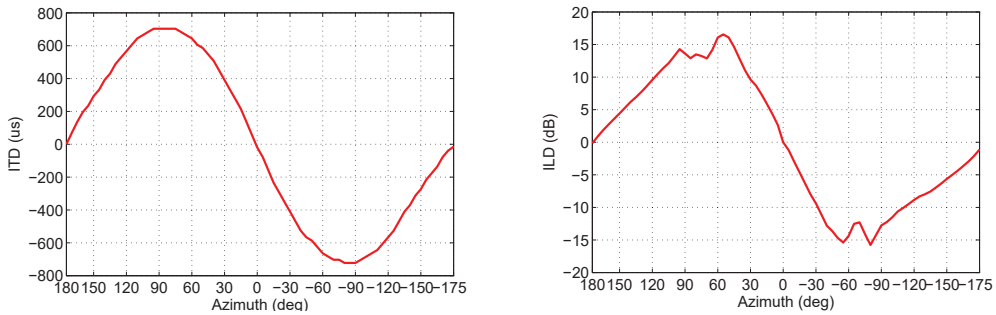
where a denotes the distance between the 2 points modeling the ear entrances, taken greater than the average head diameter.

The frequency domain version of the ITD is the *interaural phase difference* (IPD). The IPD ϕ is related to the ITD and the frequency f via the following formula [127]:

$$\phi(\theta, f) = 2\pi f \delta(\theta). \tag{A.3}$$

In addition to the delay between the 2 ears, the head is also at the root of the so-called *head shadow effect*. In fact, when the wavelengths are shorter than the head dimensions (i.e. above 1.5 kHz), a diffraction phenomenon occurs. This yields a difference of SPLs between both ears. That is, the signal at the ipsilateral ear is louder than the one at the contralateral ear. This is called the *interaural level difference* (ILD), which is usually expressed in dB. However, at LFs, the head has no effect on the sound waves because the wavelengths are quite a bit greater than the head dimensions. Figure A.10 shows the ITD and ILD as measured on a *knowles electronic manikin for acoustic research* (KEMAR) in the anechoic room of the *signal processing laboratory* (LTS) of EPFL. The sinusoid shape of the ITD is clearly apparent. Both the ITD and ILD increase with the azimuth to reach their maximum on the sides.

In 1907, Lord Rayleigh defined the *duplex theory* as the fact that the AS uses the ITD below about 1.5 kHz and the ILD above. This theory is well admitted but incomplete, since it is now known that time differences are also used in the HFs. Instead of inferring the interaural delay



(a) (b)

Figure A.10 – ITD (A) and ILD (B) measured in an anechoic room, as a function of the incidence angle θ. Positive angles correspond to the left side.

from the TFS, the AS resorts to the *interaural envelope difference* (IED). This is particularly useful when there is no LF in the targeted signal [172, Chap. 7].

Monaural cues are essentially the consequence of the pinna filtering. As mentioned in part A.1.1, the pinna is involved in the process of sound localization. In fact, for frequencies above 6 kHz, the wavelength is sufficiently short to interact strongly with the shapes of the pinna. This creates a filtering process that depends on the incidence direction of the sound source. It can be clearly evidenced when looking at the magnitude of the *head-related transfer function* (HRTF) (see e.g. [250, page 72]). The HRTF is defined in [35] as "*a specific individual's left or right ear far-field response, as measured from a specific point in the free field to a specific point in the ear canal*". A pair of HRTFs encapsulates the 3 main localization cues (ITD, ILD, and monaural spectral cues). The HRTFs are unique for everyone. They are known by the CAS, which is able to match a certain filtering to a specific direction in space.

The HRTFs of a subject can be measured in an anechoic chamber that provides free-field conditions. A sound source is rotated around the subject, who is wearing a microphone in the ear canals. The HRTF is derived by computing the TF between the signal emitted by the source and the signal captured by the microphone. The HRTF that is obtained is the combination of 2 different TFs [35, 151]: the *directional transfer function* (DTF) that contains all the spatial information, and the *common transfer function* (CTF) that is directionally independent. The latter is common to all HRTFs, and represents the information related to the measurement hardware and ear canal resonance. The CTF is usually suppressed by subtracting the mean across all HRTFs [121], and the remaining DTF is improperly called the HRTF.

ITD and ILD are used to locate sounds in the FHP. The ITD and ILD functions (Figure A.10) exhibit a front/back ambiguity. The directional filtering of the pinna helps solve this issue, because the stimuli coming from the back are more attenuated and filtered in a different manner. Furthermore, dynamic cues, resulting from head and torso motions solve the ambiguity as well. Comparing the variations of ITD and ILD resulting from slight motions of the head, the

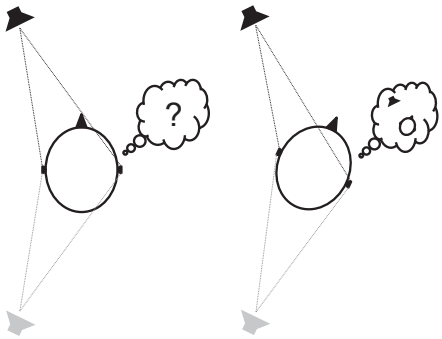


Figure A.11 – Principle of dynamic cues to solve front/back ambiguity.

AS can deduce whether the sound is on the front or in the back [230], as illustrated on Figure A.11.

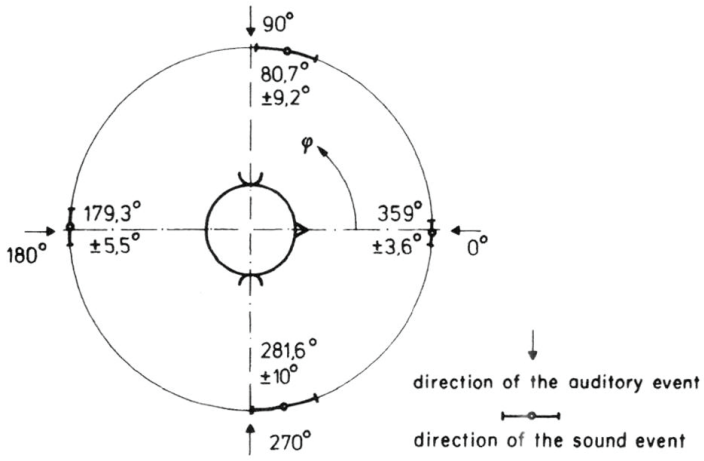


Figure A.12 – MMA measured on a large number of subjects for front/back/side azimuth. From [22, page 41].

The *minimum audible angle* (MMA) is defined as "*the smallest detectable change in angular position, relative to the subject*" [172, Chap. 7]. It serves to determine the resolution of the AS as a function of the azimuth. Figure A.10 (left panel) shows that the sinusoidal shape of the ITD provides greater variations for frontal angles than lateral angles. As a consequence, the angular resolution of the AS is better on the front than on the sides. Blauert [22, Chap. 2.1] reports the MMA measured with white noise pulses on 600-900 NH subjects, for a sound source located on the front, back and sides (Figure A.12). It can be noticed that the angular resolution is 3 times better in the front compared to the sides. In every-day life, listeners commonly circumvent this issue by turning the head until the source sound is on the front

[250, Chap. 12].

When considering distance estimation, the AS is known to present poor performance [172, Chap. 7]. Very short distances (i.e. conditions of close field) are inferred thanks to the emphasis of ILD that have a larger range comparing to the far field. The SPL changes that occur with further distances help determine distance, but only when the sound source is familiar. This is because the CAS has learnt a reference under the form of a SPL/distance ratio. Concerning familiar sounds as well, the AS is helped by the air absorption that modifies the shape of a known sound spectrum. Gardner [76] reports that it is more difficult for listeners to estimate the distance in an anechoic environment than in a reverberant area. The underlying reason is that the AS also uses the *direct-to-reverberant ratio* (DRR) to decide whether the sound is near or far.

Real environments are characterized by the occurrence of reverberation and possible interfering noises. This makes the localization more difficult. Focusing on reverberation, the direct sound of the source always reaches the ears first, then the reflections coming from various directions follow. The *precedence effect* is a prominent functionality of the AS that guarantees a relatively precise localization in reverberant surroundings. The basic principle is that the earlier arriving signal predominates over the later-arrival signals [140]. This is called the fusion phenomenon. It occurs as long as the delay between the direct sound and the first reflections is less than around 5 ms (echo threshold). After the echo threshold, more than one separate auditory event is perceived and the localization is then considerably more confused. Despite this, reverberation tends to degrade the acoustic signal, making the extraction of localization cues harder. In fact, it alters the envelope shape of the signals and modifies the modulation depth (shallowing the flanks and filling in the dips) [169]. ITD and ILD are degraded by the reflected energy reaching the ears [105]. Thus, the ILD range is decreased and possibly biased towards zero. Reverberation also decreases the *interaural coherence* (IC), which complicates the ITD extraction [105]. Rakerd and Hartmann [197] show that the localization resolution decreases from 3° to 10° when the IC goes from 0.8 to 0.3. Overall, it seems that the AS does not weight ITD and ILD in an efficient and optimal way, sometimes resorting to the worst cues [105, 197].

The ability to discriminate sounds in a noisy acoustic environment based on the source spatial separation is referred to as the *cocktail party effect*. It is particularly powerful when attempting to localize a sound of interest. Hawley *et al.* [96] actually report that the localization of the speech of interest in 3 other competing speech signals is quite good, as soon as binaural hearing is available. This contributes to a better speech understanding, and reveals the inherent relation between speaker localization and speech intelligibility.

A.2.2 Speech intelligibility

Speech signal is one of the most frequent complex sounds treated by the human AS. The mechanisms of speech production are located in 3 body regions. The subglottal system is

composed of the lungs and the vocal trachea. It delivers the sound pressure. The larynx accommodates a pair of vocal chords that vibrate in reaction to the excitation from the subglottal system. Finally, the supralaryngeal vocal tract is constituted of the pharynx, and the oral and nasal cavities. This last region modifies the vibration coming from the vocal chords. The energy of the human voice is essentially located in a bandwidth of 8 kHz. The dynamic range of speech signal is on the order of 30 dB. The vibration rate of the vocal chords is called the fundamental frequency and determines the *pitch*, which covers a frequency range between 75 and 500 Hz. The typical male speech presents a pitch between 75 and 175 Hz, while the female pitch is usually comprised between 175 and 300 Hz. Children present a pitch from 300 Hz to 500 Hz [85, Chap. 1]. Then, the vocal tract transforms the vibration of the vocal chords so as to produce some voiced or unvoiced *phonemes*, which are the basic speech components. The spectrum of voiced sounds (especially the vowels) is characterized by LF peaks that are the *formants*. These latter result from the vocal tract resonances. Different vocal tract shapes yield different resonances that produce various formants. Consonants arise from some kinds of obstructions of the vocal tract. They are divided in 4 categories: the plosives, glides, nasals and fricatives.

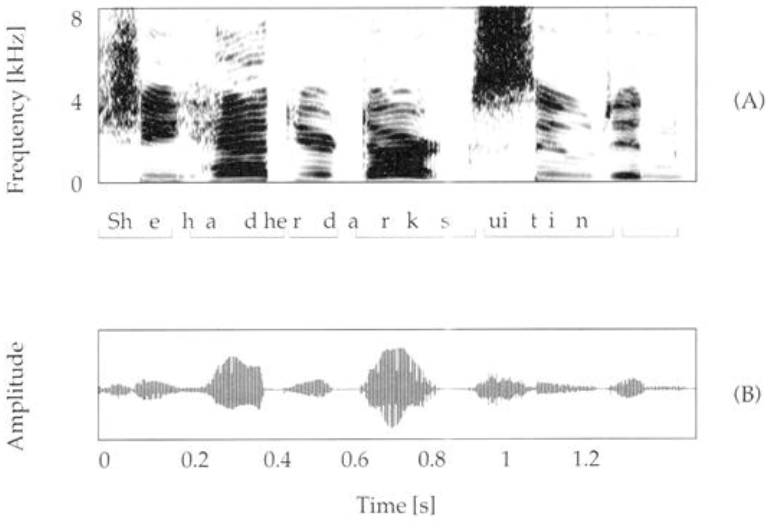


Figure A.13 – Spectrogram (A) and waveform (envelope + TFS) (B) of the sentence “She had her dark suiting”. From [85, page 77].

Speech signal can be plotted as a waveform, as shown on Figure A.13B. This figure represents the waveform of the sentence “She had her dark suiting”. The primary information shown on this figure is the envelope, a prominent component of speech signals. Indeed, the envelope brings a considerable contribution to speech understanding. When looking at the spectrum of this envelope (Figure A.14), the essential information is located below 25 Hz [85, Chap. 2]. It has to be linked with the fact that the ear temporal resolution is the best for modulation rate below 16 Hz (see part A.1). A peak can be noticed at 4 Hz. It corresponds to the average syllabic rate. The TFS is modulated by the envelope, and contains information about pitch and syllabus differentiation.

Figure A.13A shows the spectrogram of the original sentence, i.e. the frequency variations over time. The waveform of the sentence is displayed on Figure A.13B. The periodic patterns correspond to vowels, showing the pitch and formants at the LFs. Note that the first and second formants largely determine the vowel identification [85, Chap. 2]. Vowels can be clearly distinguished from consonants that present a non-periodic pattern and have the major part of their energy in the HFs. Consonants play a major role in speech understanding, since they allow to differentiate words. For instance, “live” and “life” are understood the same if one cannot distinguish consonants. Speech also contains much information such as the linguistic message, some speaker-dependent features, the environment in which the speaker is... That is the reason why speech processing is essential for communication abilities.

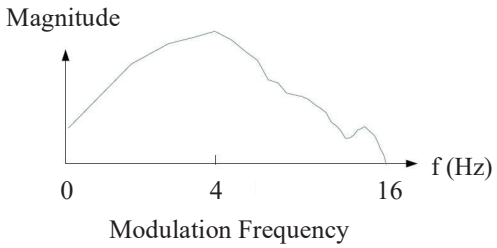


Figure A.14 – Modulation transfer function of a typical speech signal. Adapted from [64, Figure 2].

Lots of past experiments have studied speech intelligibility in reverberation and/or noise, and the effect of monaural vs binaural listening. A common way of assessing intelligibility is to determine the *speech reception threshold* (SRT), which corresponds to the SNR that yields a value of 50 % of the *speech recognition score* (SRS). The SRS is expressed as a percentage of correctly understood words in a succession of sentences or isolated words. The measurement of the SRT or SRS highly depends on the test conditions (masker and speech signals, environments...). Hence, a direct comparisons of the SRT/SRS between studies that to do not use the exact same protocol makes no sense [216].

By simulating an environment with competing speakers spatially distributed in space, Pollack and Pickett [192] report some significant differences when comparing the determined SRTs from better-ear-monaural and from binaural hearing. Binaural rendering leads to a reduction of the SRT of 5.5 dB in the presence of 7 competing speakers. It falls to 12 dB when only 1 competing speaker is present. For the same kind of interfering noises, Hawley *et al.* [96] notice an improvement of the SRS up to about 40 % in the presence of 3 competing speakers. In [30], the SRT shift equals 3 dB.

Steady state noise is known to bring more disturbance than modulated noise such as competing babble. In fact, it suppresses the *masking release*, i.e. the ability of the AS to extract information in the gaps of the masker [72]. Bradley *et al.* [25] study the combined effect of interfering noise and reverberation on the speech perception. Their results suggest that

noise is more harmful than reverberation (range of reverberation time from 0.56 to 1.95 s). Since reverberation decreases the IC, Ericson and McKinley [66] measure the intelligibility as a function of noise, for which the IC varies in successive increasing levels (0.3 to 1). They show that speech understanding is better when the noise provides a high IC value.

There exist several techniques used by the AS to preserve speech intelligibility at most in adverse listening conditions. When competing speakers are present, both frequency and temporal masking occurs. Figure A.15 depicts the spectrum and formants of a certain vowel (A), and the same vowel in a babble noise (C), of which the spectrum is shown on (B). As it can be seen, the noise affects the targeted speech spectral content, and thus the relative amplitude of the formants. Moreover, the babble noise brings additional formants. However, the original formants are still present in this case. The frequency resolution of the ear enables an efficient extraction of the first formants in the LFs, where their maximum of energy is located [85, Chap. 5]. As mentioned in [108], differences of pitch are also exploited to discriminate speakers. In other kinds of background noises, the AS resorts to a periodicity detection and harmonics identification in the TFS so as to extract the speech of interest. Considering reverberation, the effects are slightly different and essentially hit the modulation spectrum. By filling the gaps and silences, reverberation acts as a low-pass filter on the envelope. The temporal masking caused by reverberation avoids the resort to masking release as well. Note that reverberation reduces the modulation of the masking noise, which augments the occurrence duration of frequency masking.

Apart from the perceptual strategies from the PAS, the CAS also plays a major role in speech understanding, bringing knowledge of linguistic and semantic context to guess missing words. These contributions from the CAS constitutes the *cognitive processing* of speech intelligibility.

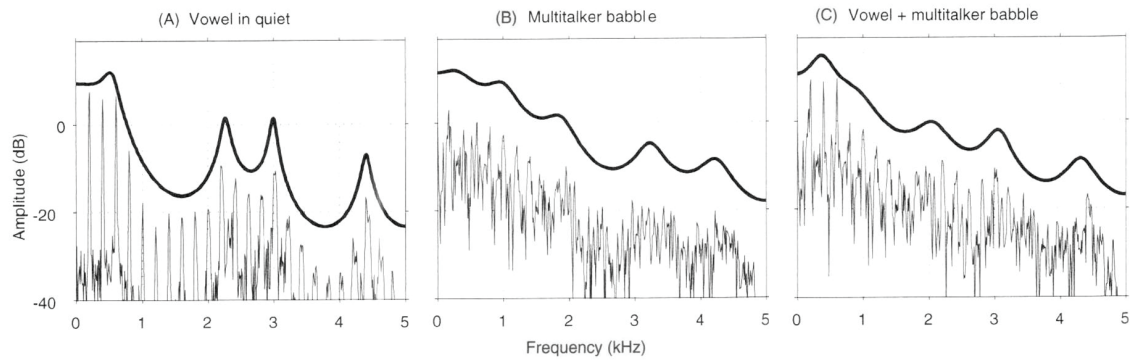


Figure A.15 – Spectrum and formants of a vowel in quiet (A), a multitalker babble composed of 2 male adults, 1 female adult and 1 child (B), and the vowel mixed with the babble noise (C). Adapted from [85, page 244].

As introduced in the previous part, binaural hearing plays a crucial role in speech intelligibility. Indeed, in real environments, the targeted speech and masker(s) are rarely located at the same position. The *spatial release from masking* (SRM) denotes "*the improvement in SRT when a*

spatial separation is introduced" between the speech and masker(s) [54]. That is, it evaluates the contribution from the cocktail party effect on speech perception. SRMs up to 8 dB have been reported in [171, Chap.7], while Greenberg *at al.* [85, Chap. 5] mention SRMs reaching 10 dB. In competing babble, spatial separation is also known to ease speaker identification [61]. Binaural hearing brings several mechanisms that help understand speech in adverse conditions. The head shadow effect is a purely physical phenomenon that naturally provides a better SNR at one of the 2 ears. The AS is able to select the ear with the best SNR in real time. That is called the *binaural switching* process. Finally, the *binaural unmasking* or *binaural squelch* denotes the ability of the AS to use the noise from the contralateral ear to improve the SNR at the ipsilateral ear. Although it does not perform a perfect noise cancelation process, it is of great help for speech intelligibility in adverse conditions.

B Appendix: Hearing impairment

The characteristics of hearing disabilities are presented in this appendix. The different types and causes of HRLs are reported. Their consequences on binaural hearing are then discussed.

B.1 Introducing hearing disorders

B.1.1 Types and origins of hearing loss

Hearing impairment can affect each component of the AS. A *conductive hearing loss* denotes "*a hearing loss due to a blockage in the outer ear, including the ear canal, or a malfunction of the middle ear*" [150]. The origins of this type of HRLs can be a cerumen blockage in the auditory canal, a damaged eardrum, stiffened ossicles, or the presence of fluid in the middle ear cavity. A reduction of audibility is the common consequence of such issues, because the stimulation of the cochlea is reduced. A HRL resulting from a damaged cochlea and/or auditory nerve is called a *sensorineural hearing loss*. The combination of both types of HRLs is a *mixed hearing loss*. Eventually, a dysfunction along the pathway between the cochlea to the brainstem is called a *retrocohlear hearing loss*.

As they result in different shapes of audiograms, conductive, sensorineural and mixed HRL can be distinguished in tonal audiometry, comparing the air and bone conduction. Figure B.1A shows the typical audiogram of a conductive HRL: air conduction (solid line) is reduced while bone conduction (dotted line), which bypasses the middle ear, is preserved. An example of sensorineural HRL is depicted on Figure B.1B. Both the air and bone conduction are reduced, as the cochlea is damaged. Finally, Figure B.1C shows the result from a mixed HRL, where the bone conduction is slightly less affected than the air conduction.

All kinds of HRLs bring about an elevation of the hearing thresholds. This can be characterized by computing the *pure-tone average* (PTA), defined as the mean of the auditory thresholds at 500 Hz, 1, 2 and 4 kHz. Depending on the value of the PTA at the better ear, one speaks about a *mild* HRL between 25-40 dB HL, a *moderate* HRL between 41-60 dB HL, a *severe* HRL between

61-80 dB HL and a *profound* HRL above 81 dB HL, according to the WHO classification [185].

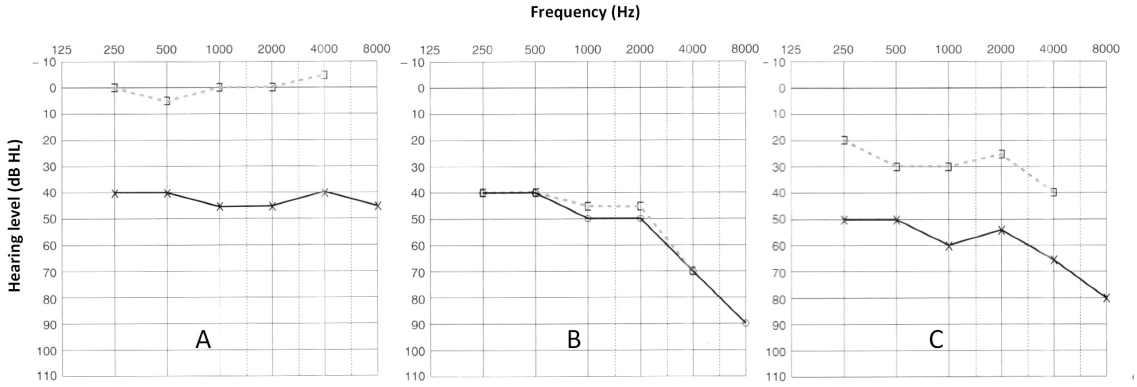


Figure B.1 – Audiogram from a conductive hearing loss (A), a sensorineural hearing loss (B) and a mixed hearing loss (C). Dotted lines are for bone conduction and solid lines are for air conduction. Adapted from [134, page 38].

There exist multiple causes of hearing disorder. The most frequent ones are an exposure to loud sounds, aging, diseases and infections, ototoxic drugs, heredity, or head injuries [250, Chap. 16]. Noise-induced HRLs are the consequence of hearing at too high SPLs for an excessive duration. Note that the loudness adaptation (Appendix A.1.2) can be a "trap", in the sense that it gives an impression of decreasing loudness and provides a form of habituation. For 200 years, the AS has been submitted to increasingly loud stimulations, due to the development of industry and new leisures (amplified music, daily-use of headphones, motorbikes...). In such a short time, the natural evolution did not manage to elaborate an efficient protection of the ears against this danger [149, Chap. 9]. Too loud sounds primarily affect the inner ear, but very impulsive stimulations (e.g. explosions) can also lead to damages of the middle ear, due to the short reactivity of the acoustic reflex. Outer hair cells are known to be more vulnerable than inner hair cells to excessive SPLs [171, Chap. 1]. Their stereocilia swell and eventually get destroyed [250, Chap. 16]. The amplification property of those cells is then lost and the AS appears to be less sensitive to sound in the corresponding frequency regions. The increase of the absolute thresholds up to 50 to 60 dB is the usual consequence of outer hair cell dysfunction. Higher HRLs result from the combined destruction of the inner and outer hair cells [171, Chap. 2].

Presbycusis, i.e. age-related HRL, denotes a disorder of the inner ear as well, although the entire AS is actually concerned [149, Chap. 11]. It is characterized by a slopping HRL, i.e. a HRL that increases with the frequency. This is because the base of the cochlea (HF processing) is quite a bit more exposed to sound stimulation than the apex (LF processing). Over time, the hair cells gradually die, which causes a progressive loss of frequency from the HFs to the LFs.

Additionally, diseases can also hurt the PAS structures. 70 % of young children are regularly infected by otitis median, usually resulting from an infection of the eustachian tube, which brings fluid into the middle ear cavity [149, Chap. 10]. A medical treatment is mandatory to

avoid any spread to the inner ear, which would cause e.g. a meningitis and probably yield a permanent hearing impairment or a total deafness.

B.1.2 Sensorineural hearing loss

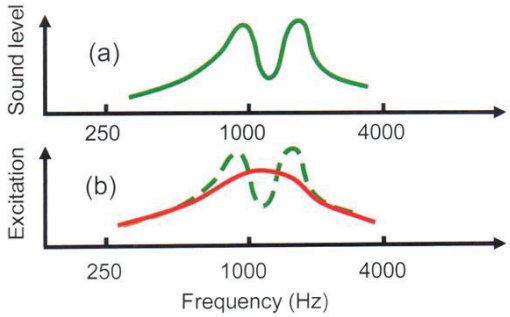


Figure B.2 – Frequency representation of a signal with spectrum (A) in the AS of a NH (green line) or HI (red solid line) subject. From [58, page 4].

This thesis primarily deals with sensorineural HRLs. The expression "hearing impaired" (HI) subject now concerns subjects encountering a sensorineural HRL only. Such a HRL is far from being a simple loss of audibility. Besides, a common complaint of subjects presenting a sensorineural HRL is that they can hear speech but cannot understand the content. Sensorineural HRLs are characterized by elevated hearing thresholds, due to the dysfunction or destruction of the hair cells. The entire disappearance of hair cells at a certain place in the basilar membrane is a *dead region*. Such a destruction leads to various issues, among which a decrease in frequency resolution, in time resolution and in loudness perception [58, Chap. 15].

As seen in part A.1.3, the frequency processing of the AS is related to the existence of auditory filters. The critical bandwidth of those filters has shown to be larger in HI subjects than in NH subjects, especially in LFs [171, Chap. 3]. This augments their sensitivity to LF maskers and diminishes their ability to take advantage of the spectrum differences between the signal of interest and the masker(s). As previously mentioned, the frequency resolution of the AS decreases with increasing SPLs. Therefore, even at high SPLs, HI listeners cannot really benefit from frequency selectivity. Figure B.2B shows a typical representation of a signal, with a spectrum depicted on Figure B.2A, in the AS of a NH subject (green dotted line) or a HI subject (red solid line). The details of the signal spectrum are significantly blurred in the HI listener's case.

The temporal resolution is also modified. The masking release is smaller than in NH subjects, and even absent when the targeted signal spectrum falls in a dead region. Additionally, the temporal integration is distorted. This leads to an abnormal perception of loudness. Indeed, HI subjects experience a disproportionate increase of loudness above their hearing threshold, contrary to NH subjects. Their elevated absolute thresholds, combined with a normal pain

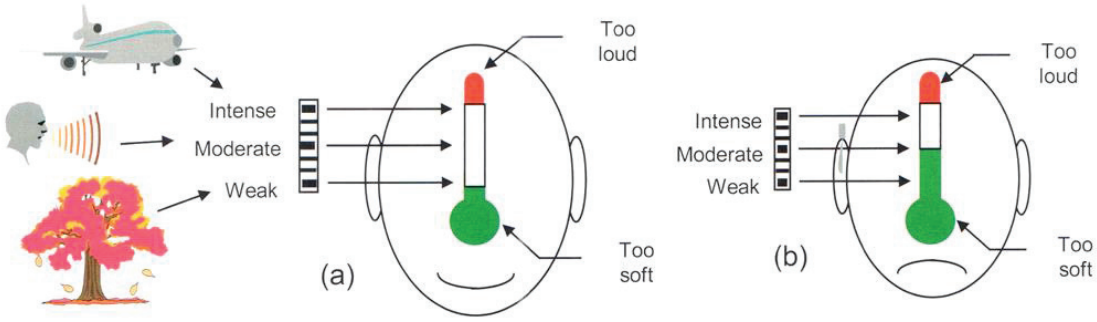


Figure B.3 – Illustration of the recruitment phenomenon. (A) shows the auditory dynamic range for a NH listener, and (B) the one for a HI listener. Adapted from [58, page 3].

threshold, cause the dynamic range of their AS to be significantly lower than NH subjects. This phenomenon is known as the *recruitment* [171, Chap. 4]. It is depicted on Figure B.3. This picture shows an example of the auditory dynamic difference between a NH subject (A) and a HI subject (B). The SPL range of too soft sounds is quite a bit larger in the HI case while the too loud zone remains the same as the NH subjects. Finally, it is prominent to mention that the AS performance of HI listeners presents an important inter-subject variability, even for a similar HRL [171, Chap. 2]. That is, subjects with an identical HRL can exhibit totally different performance.

B.2 Hearing impairment and binaural hearing

B.2.1 Localization

Part A.2 highlighted the importance of cochlear processing in binaural hearing. It would be expected that HI subjects present lower performance than NH listeners in localization tasks. The perception of the ITD, IED and ILD have effectively shown to be less precise. Some studies reported in [171, Chap. 7] indicate that the detection of ITD variations at 500 Hz differs from 25 μs for NH subjects to 210 μs for HI listeners. The IED variation detection goes from 25 μs up to 250 μs at 4 kHz, where the detection of ILD changes increases up to 3 dB. HI listeners suffering from an asymmetrical HRL (i.e. a unilateral HRL or different degrees of HRLs between the 2 ears) exhibit similar performance in the detection of ITD, but a quite a bit lower sensitivity to ILD changes.

Overall, HI subjects are better in binaural cue detection for broadband stimuli than for narrow-band ones [61]. This is presumably because of their degraded frequency resolution [179]. In addition to their elevated hearing thresholds, the poorer performance in terms of ITD detection might be due to some changes in the propagation time of fluids in the cochlea [171, Chap. 7], while deteriorations of the ILD perception would be the consequence of loudness distortion (temporal integration), more marked for asymmetrical subjects. HI listeners resort to monaural cues quite less than NH subjects. This is explained by their limited access to HFs

and their worse frequency selectivity. It prevents the perception of the peaks and notches of the HRTF. As for the precedence effect, it has been shown to be significantly less efficient in HI listeners, leading to pronounced difficulty in localizing sources in reverberant surroundings. The origin of this issue would be related to the impaired temporal processing.

It is noteworthy that HI subjects do not usually pay attention to their localization ability. They become conscious of their relative difficulties in localizing sounds in space when they are asked to [24]. Curiously, the localization performance of such subjects appears to be minor in the horizontal plane, at an adapted SPL, despite a high individual dependence [171, Chap. 7]. This is not the case for asymmetrical listeners. Larger MMAs were reported in lateral azimuths for all HI listeners, and especially on the side of the damaged ear for asymmetrical subjects. Abel *et al.* [1] indicate a decrease of 15 % of localization accuracy between young adults and moderate presbycusis subjects. Furthermore, the underuse of pinna cues in HI subjects yields a higher rate of front/back confusions. The reduced frequency resolution leads to persistent issues in localizing sound sources in noise [61]. Despite a detrimental effect of hearing disorders on localization cue extraction, the CAS acclimatizes to this and somehow preserves a good accuracy in localization tasks. The continuous training that takes place by matching visual and auditory events may be the main reason for that [24].

B.2.2 Intelligibility

It is a common complaint from HI subjects to understand almost nothing of speech in noisy places, whereas quiet areas yield a pseudo-normal intelligibility. Yet, this is not valid for HI subjects exhibiting dead regions. Such subjects present high SRTs in noiseless conditions as well, even if audibility is restored [215]. A substantial proportion of HI subjects often combines the acoustic speech stimulus with lip reading, and even sign languages in case of profound HRLs. Several disorders in the AS are at the root of such difficulties. As described in Appendix A.2.2, NH subjects resort to various techniques to keep a good perception of speech in the presence of reverberation and/or background noise. This includes the TFS processing, the masking release and the SRM. All those features are significantly lessened in HI listeners.

Due to their degraded frequency selectivity, HI subjects experience difficulties when processing the TFS of speech signals. As a consequence, their performance in differentiating pitch and extracting formants in noise is worse than NH listeners [171, Chap. 7]. That is why they put a greater emphasis on envelope processing, where they perform almost as good as NH subjects [220]. Vowels, which present a broadband spectrum, are easier to understand than consonants, especially the ones that have a narrow band spectrum, e.g. the fricatives. This is highly problematic, since it has been recalled in Appendix A.2.2 that consonant identification plays a major role in speech perception. Also, female voices are usually less intelligible and appreciated by HI listeners. This is because the harmonics of their formants are higher in frequency compared to the male speech. Note that TFS sensitivity naturally decreases with age, independently of the presence of HRLs [80, 98]. This would be explained by some CAS

disruptions [59, 72].

The loss of temporal resolution is another factor that accounts for a great part of intelligibility concerns. In fact, benefits from masking release (listening in the dips) are extremely weak in HI subjects, due to the occurrence of an unavoidable temporal masking [80, 83]. This is demonstrated by Festen and Plomp [72] that report no difference in intelligibility scores whatever the nature of the masker (steady state or modulated). Smits and Festen [216] make SRT measurements in a group of NH subjects and a group of HI subjects, with a steady state noise. The difference of SRT between the 2 groups comes to 5 dB. Then, they repeat their measurements using a fluctuating masker [217] and study the difference between steady state vs fluctuating noise in both groups. The difference reaches 6 dB for the NH subjects, while it is only of 1.5 dB for the HI group. This confirms the conclusion of Festen and Plomp about the absence of an efficient masking release in HI listeners. Increased temporal masking is also problematic when louder vowels mask the following softer consonants, making it even more arduous to distinguish consonants. Note that, similarly to the frequency selectivity, the temporal resolution naturally decreases with age.

Some benefits from binaural hearing are absent in HI listeners as well, and they do not disappear with age, i.e. old subjects without HRL present the same performance of spatial processing as young NH subjects [83]. HI listeners are shown to take a poor advantage of spatial separations between the signal of interest and the interfering noise(s) [180]. Indeed, Moore [171, Chap. 7] reviews studies that report SRMs of 10 dB for NH listeners, whereas they only reach 4 dB for HI listeners. This is mainly due to the low audibility of HFs, where the essential part of the head shadow effect takes place. Nevertheless, Best *et al.* [21] showed that there is no evidence indicating that the binaural switching would be less efficient in HI subjects.

Despite their relatively good performance in binaural localization, HI listeners suffer from large difficulties in understanding speech, especially in complex environments. That is why they require some help, which is primarily provided by HAs.

C Appendix: Simplex optimization method applied to the localization algorithm

In order to optimize the performance of the BLA (see Chapter 3.1.3), some first optimization trials were done with the *simplex optimization method*, described in [143]. Its principle is to quickly converge to a local minimum of the function g. As explained by Lundstedt *et al.*, "*a simplex is a geometric figure with* $(L+1)$ *corners where p is equal to the number of variables in a L-dimensional experimental domain*". In the current case, C is equal to 3, and the simplex is thus a tetrahedron. The processus is the following:

1. Start with a given point in the $(\xi \times \lambda \times \rho)$ space, and compute the coordinates of 3 additional points, so that the combination of them forms a regular tetrahedron with a defined edge d_{def},

2. Compute the distance D associated to the 4 vertices of the tetrahedron,

3. Mirror the vertex that yields the greatest distance D (i.e. the worst performance) through the barycenter of the other corners. This creates a new simplex,

4. Repeat the last operation until the response does not improve any longer,

5. The vertex of the last tetrahedron leading to the shortest distance is a local optimal point.

This process can also be described in a mathematical way. Given the 4 vertices of the first simplex V_{best} (corner with the smaller distance), V_1, V_2 and V_{worst} (corner with the bigger distance), the distance d_{best} between the barycenter μ of the triangle formed by V_{best}, V_1 and V_2, and the best vertex V_{best} is computed as follows:

$$d_{\text{best}} = \sqrt{\left(\mu_1 - \xi_{\text{best}}\right)^2 + (\mu_2 - \lambda_{\text{best}})^2 + (\mu_3 - \rho_{\text{best}})^2}, \tag{C.1}$$

where μ_1, μ_2 and μ_3, are the coordinates of the barycenter μ. The distance d_{new} between the new point V_{new} replacing V_{worst} and μ is:

$$d_{\text{new}} = \sqrt{d_{\text{def}}^2 - d_{\text{best}}^2}. \tag{C.2}$$

Note that the argument is always positive, because $d_{\text{def}} > d_{\text{best}}$ by definition.

Then, the coordinates of V_{new} can be calculated:

$$Y = \begin{cases} \xi_{\text{new}} = \mu_1 + Y(1)\, d_{\text{new}} \\ \lambda_{\text{new}} = \mu_2 + Y(2)\, d_{\text{new}} \\ \rho_{\text{new}} = \mu_3 + Y(3)\, d_{\text{new}} \end{cases}, \tag{C.3}$$

with Y being the vectorial product between $\overrightarrow{V_{\text{best}} V_1}$ and $\overrightarrow{V_{\text{best}} V_2}$.

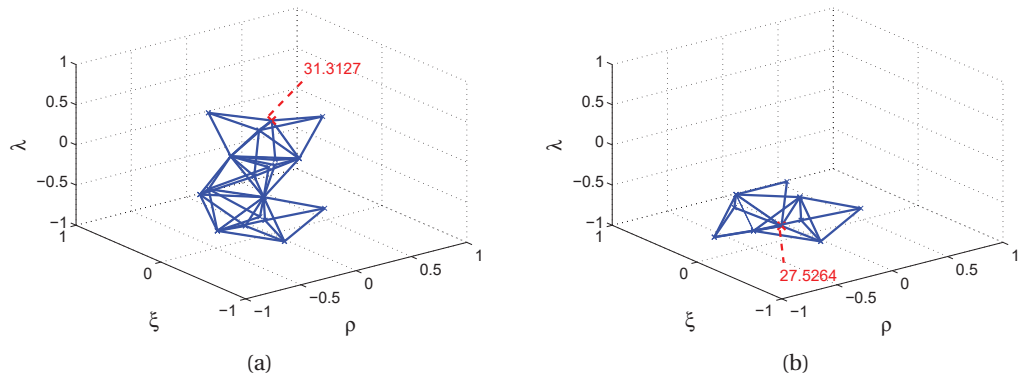


Figure C.1 – Results from the simplex optimization with the same dataset (male speech in the classroom) for a certain starting tetrahedron (A) and a different starting tetrahedron (B). The final optimal point is given in red, and the associated distance D is displayed.

Note that one must take care that the condition $|\xi_{\text{new}}| < 1$ and $|\lambda_{\text{new}}| < 1$ and $|\rho_{\text{new}} < 1|$ is respected. Otherwise, the new point is out of the range and has to be removed. In this case, the procedure of the determination of V_{new} is applied to the second point of the tetrahedron yielding the worst performance.

Unfortunately, the simplex minimization method has not been really convincing. Indeed, it appears to strongly depend on the initial chosen tetrahedron, so that different local minima $(\xi_{\text{OPT}}, \lambda_{\text{OPT}}, \rho_{\text{OPT}})$, exhibiting completely different performance, have been output by the procedure. Figure C.1 shows an example of this issue, displaying the outcomes from the

simplex optimization procedure using the same dataset (male speech in the classroom) and 2 different starting tetrahedrons. As shown, 2 different optimal point are found, with different distance values (D = 31.31 on Figure C.1A, and D = 27.53 on Figure C.1B). In this case, one would prefer the second combination. However, one can also legitimately imagine that some even smaller values of D could be reached, with other starting tetrahedrons. Thus, it makes sense to try a different method of optimization, namely the response surface design (Chapter 3.1.3).

D Appendix: Algorithm implementation on the embedded prototype

Here is detailed the way how the BLA and the BSA have been integrated to the prototype. In particular, the conversion from a floating- to a fixed-point resolution is discussed.

D.1 Hardware

The specifications of the hardware embedded in the prototype have been detailed in Chapter 1.4.3. The prototype is depicted on Figure D.1. It is composed of the following elements:

- 2 BTE HAs,
- 2 RF receivers plugged on the DAI of the HAs. The RSSI and the audio signals from the HA microphones are transmitted to the BWU via wires,
- The BWU that contains the electronics of the system, especially the microprocessor Atmel ARM9. It gets as input the wireless audio from the emitter, the audio signals captured by the HAs and the RSSI at both sides. Once the spatialization is applied, the binaural audio is sent back to the HAs through the wires.

The clock of the microprocessor is fixed to 19.2 MHz [182]. The frequency response of the BWU is flat between 100 Hz and 1 kHz, then slowly decreases by 10 dB up to 8 kHz. The overall delay (buffering / analog-to-digital conversion / ARM9 / digital-to-analog conversion) is around 16 ms.

D.2 Software implementation

D.2.1 Floating- to fixed-point conversion

The development of the BLA and BSA has been conducted under Matlab and Simulink, which operate with a floating-point resolution. On the contrary, the digital processing in the BWU is performed with a fixed-point resolution of 32 bits. Note that, when possible, one prefers to

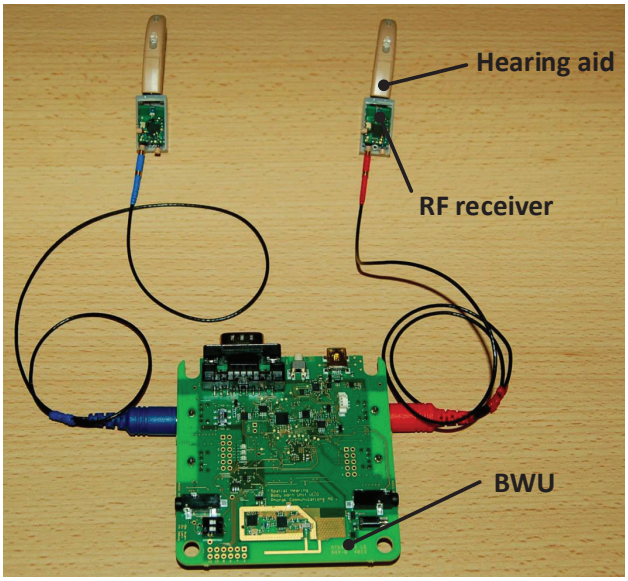


Figure D.1 – The hardware embedded in the prototype.

resort to 16 bits only. It means that all the decimal variables of the codes are truncated when integrated in the prototype. This may cause some problems for the good reproduction of the algorithm performance provided by the prototype. In fact, the rounding of the values can change the behavior of the blocks. For instance, a modified value of the sine error derived in the IPD block (Equation 2.16), coming from previously rounded signals and computation results, may make certain adverse frames being processed by the BLA. This would yield a wrong localization of the speaker that does not occur when the algorithm is running under Simulink. Also, IIR filters are widely used at various stages of the algorithm, although they are known to be sensitive to truncations. Therefore, the rounding error can propagate and result in an unstable output. Before integrating the code on the ARM9, one has to make sure that the fixed-point resolution does not disturb the processing.

The Matlab Coder toolbox [152] enables to automatically generate C-code from a code written in Matlab. Knowing the range of all variables, the users can define the resolution of them prior to that conversion. Then, they can check if the fixed-point translation does not have any consequence on the outcomes. Thus, one can choose how many bits are devoted to a given variable, and what must be its fraction length. Figure D.2 gives an example of that conversion from floating point to fixed point, for the ILD block. The variable called "AUL" correspond to a 128-sample frame of s_L (highlighted in red). This variable is known to take values between -1 and 1 (full scale resolution at the output of the analog-to-digital converter). A fraction length of 15 bits can thus be devoted to the decimal part, the last bit coding the sign. The variable *Denum*, highlighted in green, stores a part of the coefficient of an IIR filter used in the ILD block. The columns "Sim Min" and "Sim Max" indicate the maximum values of *Denum*. 1 bit is taken for the sign, and 3 bits are required to code the integer part up to 4. The remaining

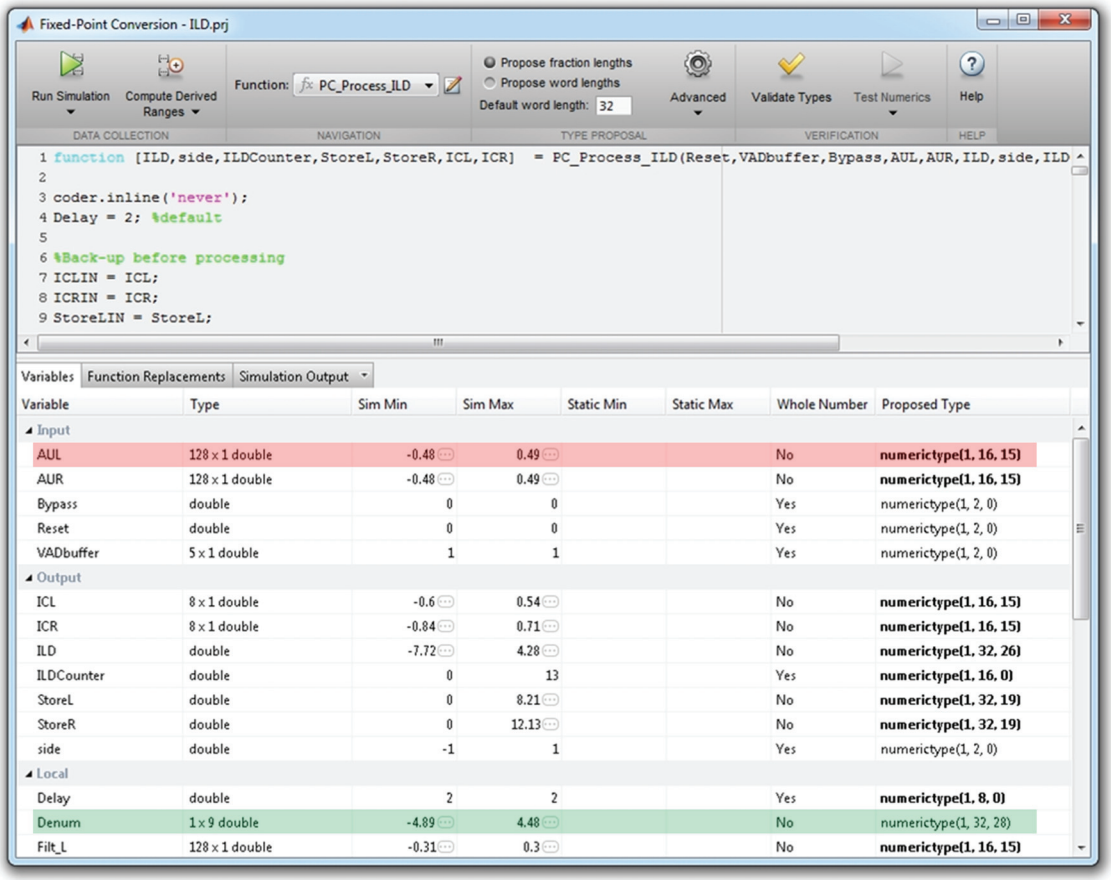


Figure D.2 – Example of the fixed-point conversion of the Matlab code corresponding to the ILD block. The variable highlighted in red and green represent an audio frame of s_L and a set of an IIR filter coefficients respectively.

bits are devoted to the decimal part. A 32-point resolution is used for *Denum*, so as to limit the risk of some possible propagation errors mentioned before.

Once the variable resolutions are defined, one has to check whether the magnitude order of the error between the fixed-point and floating-point processing is satisfactory. An example of the output of the RSSID block is depicted on Figure D.3, showing the floating-point RSSID (Figure D.3A), the fixed-point RSSID (Figure D.3B) and the error between the 2 (Figure D.3C). The bit repartition has been set so that the truncation error is smaller than 0.002 dB, which is perfect and cannot result in any false RSSID value.

The adequate conversion of all blocks in a fixed-point resolution makes possible the generation of the corresponding C-code. This allows the integration of the BLA and BSA in the hardware.

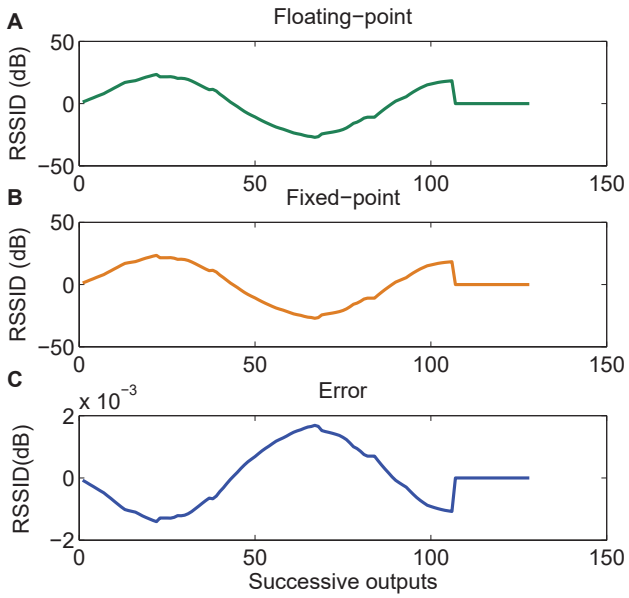


Figure D.3 – The floating-point (A) and the fixed-point (B) RSSIDs, and the error (C) between them.

D.2.2 Integration of the algorithms

Once the process is implemented in the BWU, the most prominent point to check is the real-time functioning of the algorithm. The whole processing have to take less than 4 ms (rate of the incoming frames), from the reception of the signals to the spatialized rendering in the HAs. Table D.1 reports the maximum values of the computation times required by each block. The IPD, ILD and Spatialization blocks are the most demanding ones. The total processing time appears to exceed the limits of the maximum allowed duration for the computations. Besides, a certain margin of time (300-500 μs) would be desirable. Note that the CE blocks is removed from the BLA because it has been impossible to reach a reasonable computation time (i.e. needs more than 2 ms).

The concatenation of two 4-ms frames for the analysis of the speech signal (128-point processing frames) help circumvent the issue of the excessive computation time. Indeed, the processing has been shared between both 64-sample buffers, where only the signal conditioning and BSA are launched at each frame. This enables to reduce the processing time down to around 2.5 ms on each 4-ms frame [11].

Considering the memory usage, the ROM required is around 40 kB [10]. In the most critical part, the RAM usage is lower than 3.4 kB. These results are in agreement with the specifications stated in Chapter 1.4.3.

Blocks	**Maximum processing time (μs/frame)**
Signal conditioning	465
VAD	450
ILD	750
RSSID	25
IPD	1250
Localization & Tracking	320
Spatialization	750
Total	**4010**

Table D.1 – Maximum computation time required by the various blocks of the entire algorithm. Taken from [11].

E Appendix: Assessing intelligibility, localization and preference

The appendix reports the extensive review of the literature that has been conducted in a view to write the protocol of the clinical trial. As shown on Figure E.1, 3 kinds of tests have been investigated: intelligibility (24 reviewed studies), localization (25 reviewed studies) and preference-rating (10 reviewed studies) tests. The aim is to get some conventional trends about the involved subjects, procedures, stimuli, and hardware. After each review of the 3 categories of tests, some ideas and possible applications are discussed.

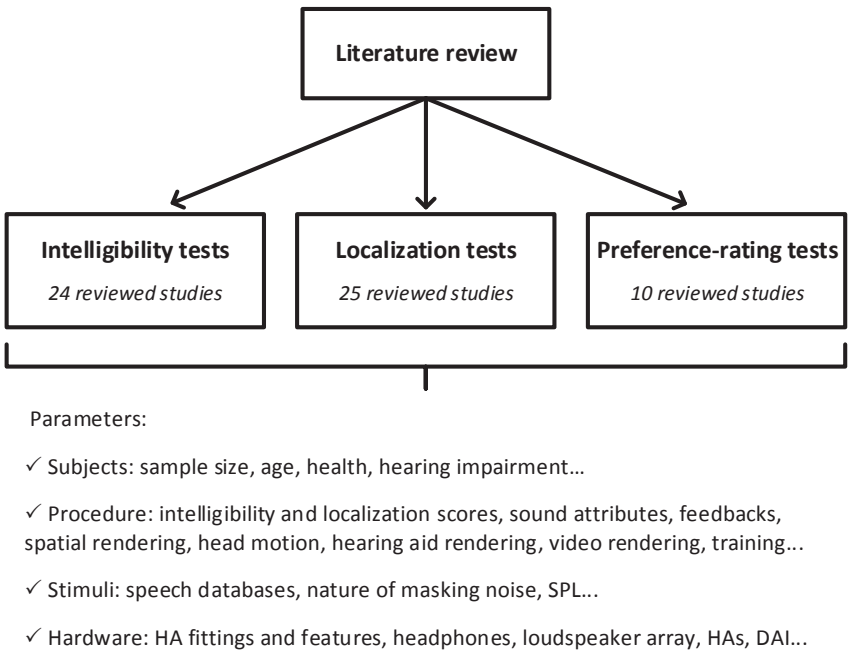


Figure E.1 – The different types of reviewed tests, with the various investigated items.

E.1 Intelligibility test

Sample size

The reported studies involve an average sample size of 19 HI subjects, with a SD of 10. The minimum size is 7 [60] and the maximum reaches 44 [136]. The required number of subjects mostly depends on the tolerated variations among listeners, in terms of HRLs and general performance. When it comes to NH listeners, the average sample size is 12.5 subjects (SD = 4.5), the minimum is 6 [7] and the maximum is 20 [62, 72]. This suggests that less NH than HI subjects are usually, presumably because they present more homogenous results. Lots of studies are working both with NH and HI participants [7, 62, 72, 80, 88]. The involvement of NH subjects is often motivated by the need for getting some reference data for comparison with HI listeners' outcomes. As an example, it is possible that no beneficial effect of the BSA appears in HI listeners, whereas there is one for NH subjects. At least, this would confirm the hypothesis that a HA is a suitable hardware for delivering spatialized sound.

Age

All the reviewed experiments involve adult subjects, often elderly. The noteworthy exception is in [229] that includes 5 children (11-15 y.o.) with 5 adults (20-55 y.o.). The main advantage of working with children is that they form a great part of the end-users of the developed system. Their hearing characteristics may differ in a significant way from adults, especially if one considers presbycusic adults. Even though the audiograms of a child and an adult are similar, some considerable performance differences can be noticeable, because the processing done by the CAS is not similar [136]. However, there are several drawbacks when working with children. As recalled by Van Hemel and Dobie [236], their cooperation, as well as the test duration, are often quite limited. Moreover, they mention that "*children with congenital or prelingually acquired hearing loss are often delayed in the development of speech and language skills and have restricted vocabularies*". There is a clear dependence on the age of the tested children: Gravel *et al.* [84] report that older children (more than about 11 y.o.) and children with a larger vocabulary are able to tolerate noisier speech than younger children. Working with adults, especially with elderly, is definitely simpler because they present more homogenous HRLs (the presbycusis pattern is most commonly a sloping HRL), they have more time (if retired) and they are easier to recruit.

Hearing and health

Considering the health of the participants, Lewis *et al.* [136] report a certain number of interesting selection criteria:

- No history of chronic or terminal illness, psychiatric disturbance, or senile dementia as reported by the participant,
- No history of stroke or cerebral vascular disorder with a paresis or aphasia as reported by the participant,

- Willing and able to give written informed consent to participate in the investigation, as noted by their signature on a consent document.

The most common required characteristics of hearing impairment is the symmetry. In almost all the reported studies, symmetry is a selection criterion to ensure a well-balanced binaural processing in the AS. There exist 2 major approaches to define and ensure symmetry:

1. A comparison of each frequency value of the audiogram between both ears independently. In this case, several thresholds have been found in the literature, and the most usual is a maximum allowed discrepancy of 15 dB [60, 104, 114, 123],

2. A comparison of the average left and right HRLs, often using the PTA. [7].

All the reviewed studies except one involve HRLs with a sensorineural origin. When it comes to the audiogram, the majority of the reported studies retains subjects with a HF sloping HRL. The following criteria by Lewis *et al.* [136] can be used:

- Ear inspection via otoscopy within normal limits,
- Normal middle ear function bilaterally (±100 dekapascals) as indicated by tympanometry,
- No evidence of conductive or retrocochlear pathology as indicated by pure-tone testing and immittance measurements,
- No air-bone gap greater than 10 dB at any test frequency as indicated by pure-tone results,
- Symmetrical HRL that does not differ by more than 15 dB at most at any audiometric test frequency as indicated by pure-tone test results.

Normal hearing is usually defined by hearing thresholds below 20 dB HL between 250 Hz and 8 kHz. It is common to add a margin and limit the maximum threshold to 15 dB HL [80, 81, 96].

E.1.1 Procedure

Intelligibility index

As explained in Appendix A.2.2, the 2 major indices to quantify speech intelligibility are the SRT and SRS. Over the reviewed studies, 12 express the intelligibility in terms of SRT and 11 resort to the SRS. Practically, the intelligibility is commonly measured by counting the number of well understood consonants, single words, sentences, or keywords in a sentence [196]. Those keywords are generally nouns, transitive and intransitive verbs, and adjectives. Nilsson *et al.* [176] indicate that the intelligibility performance obtained with the SRS is reliable but also

inherently limited by the floor and ceiling effects (i.e. getting only some 0% or 100% scores), which underlines the importance to have a large number of conditions (that is, estimate the SRS for various SNRs). This is to cover the range between 0 and 100% of intelligibility. On the other hand, the main advantage of the SRT is that it is not concerned by any floor or ceiling issue, since the procedure adapts to subjects.

The procedure to estimate the SRT is well established and is available in [191]. Conversely, there is no conventional approach to get the SRS. Here are some examples reported in the literature:

- Bradley *et al.* [25] test some NH subjects with 4 SNRs (-5, 0, +5, +10 dB) varying the noise level,
- Ericson and McKinley [66] test some NH subjects with 6 SNRs (-47, -37, -32, -22, -12, +8 dB) varying the noise level,
- Thibodeau [229] tests HI subjects with 5 SNRs (+4, +11, +16, +21, +30 dB) varying the noise level,
- Köbler and Rosenhal [123] test some HI subjects with 1 SNR (+4 dB), with a speech level fixed to 73 dB SPL,
- Picou *et al.* [187] test some HI subjects with 4 SNRs (+3, +6, +9, +12 dB), with varying noise and speech levels. All subjects are not tested with the same SNRs, so as to avoid the ceiling and floor effects. All subjects start with +9 dB. If performance is close to 100%, the SNRs tested are +3 and +6 dB. If performance is close to 0%, the SNRs tested are +9 and +12 dB. Otherwise the SNRs tested are +6 and +9 dB.

As expected, one can notice that the SNRs tested for NH and HI subjects are different, even when HI listeners are wearing their HAs.

There are different approaches to grade the intelligibility of a speech stimulus. In the case of single words, the method is straightforward as the answer can only be true or false. When using sentences, the question is whether the same weight must be given to all words contained in the phrases. Some studies require the sentence to be repeated without a single error, i.e. marked either 0% or 100% [27, 61], while others are counting the number of keywords correctly understood [96, 123, 187]. Thibodeau [229] gives 3 indicators, which are the number of times the first word is correctly identified, the number of times the whole sentence is correctly identified, and the number of words correctly identified for each sentence. It is also possible to present to subjects several sentences, and ask them to choose which one they have heard [60, 88]. A training phase of a few sentences is generally included at the beginning of the test.

Spatialization and hearing aids rendering

One of the aims of the clinical trial is to determine whether the use of spatialized speech does not degrade the speech perception compared to a diotic speech. Therefore, there is the need to

compare the SRS obtained with some spatialized and diotic stimuli for the same test conditions. Many studies use headphones to conduct their investigations, either with monaural speech signals [61, 72, 80, 81, 215, 220], or with spatialized speech signals [54, 88, 96, 251]. A few studies also resort to both diotic and spatialized stimuli. In this case, one has to ensure equal loudness in both conditions, to allow some valuable comparisons of the outcomes. While it is easy to measure the SPL of a speech stimulus in a diotic situation (i.e. measurement of the level at only one ear using a manikin), it is more complex for spatialized stimuli, where both ears do not have access to the same SPL. Obviously, a simple addition of the left and right ear levels makes no sense. Here are 2 procedures suggested in the literature:

- Begault and Erbe [15] measure the long-term RMS value of a speech-spectrum noise at one ear, either with the diotic or with the 0°-spatialized stimulus. Then, the levels are equalized. The same gain is applied to all the other spatialized directions, assuming that only the natural filtering of the body changes the loudness between the different locations. This is true if all HRTFs come from the same database (i.e. measured in the exact same conditions).
- Drullman and Bronkhorst [61] fix a certain loudness for the monaural condition. In the binaural condition (speech spatialized at 45°), only the level at the ipsilateral ear is measured, and the stimulus one is set so as to provide the same loudness.

Dealing with HA rendering, no conventional procedure to measure intelligibility with aided subjects has been reported in the literature so far. Over 8 reviewed studies investigating speech perception through HAs, 6 are using a loudspeaker or an array of loudspeakers to play the stimuli, and only 2 [215, 229] are resorting to the DAI.

E.1.2 Stimuli

Speech

There are numerous databases of speech content available for testing the intelligibility in English and Dutch (VU test material, IEEE sentences, Harvard sentences, VCV, Bamfort-Kowal-Bench sentences...). The language of the proposed clinical trial is French, as it takes place in the French-speaking part of Switzerland. 2 major databases exist in this language:

- The *hearing in noise test* (HINT) database, from the Collège National d'Audioprothèse (Paris, France) [57]. It consists in 5 lists of 20 simple meaningful and phonetically-balanced sentences (e.g. "*La fille lave ses mains*", "*Le chien ramène le jouet*") with 4 to 7 words,
- The *semantically unpredictable sentence* (SUS) database developed by Raake and Katz [196], which is made of 288 semantically unpredictable (i.e. meaningless) sentences, arranged in 24 lists of 12 sentences (e.g. "Le chien lutte sous la plage rouge", "la robe sourde voit l'ours").

No significant effect of the speaker gender has been reported in the literature. Both the HINT and SUS databases are available for a male talker.

Over the 24 reviewed studies, 14 are evaluating sentences, 7 are using only words and 3 evaluate other contents, such as syllables. There are more tests involving sentences, presumably because it is closer to some real situations. In most cases, the content consists in some everyday short sentences, with a few syllables and keywords. Raake and Katz [196] highlight the prominence that all sentences and lists of sentences are similarly intelligible. This is the primary reason why one should use a phonetically-balanced dedicated database. The authors also indicate that, as long as no translation is intended in other languages, it is better to employ a large lexicon and more conversation-like topics, rather than a limited lexicon. This is to reduce the risk of some possible training effects.

The question of whether the sentences should have a meaningful content or not is an open point. Meaningful sentences are more predictable because they allow to deduce missing words via a cognitive processing. Thus, it is possible that a sentence is repeated correctly thanks to a kind of guessing, rather than with a purely speech understanding. As mentioned by Benoît *et al.*[18], "*Meaningful sentences provide semantic and syntactic contextual cues whose effect on intelligibility scores cannot readily be quantified. This makes it virtually impossible to construct lists of sentences which are balanced in terms of their complexity*". The use of meaningful sentences is common in the literature and presents the advantage of sounding natural to subjects, while meaningless ones can be disturbing.

Noise

Various kinds of masking noises are used in the literature. The most common ones are the *speech-shaped noise* (10 studies over 24) and the babble noise (5 studies over 24). The speech-shaped noise corresponds to a white noise that has been filtered so that its long-term spectrum is the same as the long-term spectrum of the speech used. Lewis *et al.* [136] indicate that this kind of masking noises is the one that has "*the most deleterious effect on speech perception*" for both NH and HI subjects. Otherwise, the babble noise (also called "cocktail party noise") is a mixture of several competing speakers. Other types of reported noises are e.g. a modulated speech-shaped noise [72, 80, 81], a time-reversed speech [72], classroom noises [229]...

SPL

The question of the SPL rendering is important. Indeed, the stimuli should be loud enough to guarantee the audibility, while not becoming harmful for the ears. Table E.1 and Table E.2 gather the playback levels for the speech and noise signals in tests involving NH and aided HI subjects respectively. It is recalled that a long-term level of 60 dB SPL corresponds to a level of about 55 dB(A), for a speech signal.

Overall, the basic levels delivered between the speech and masker signals are similar, for both NH and HI listeners. The difference of sound strength between both categories is fairly minor

(around 5 dB more for the HI subjects, for both voice and noise). The underlying reason is that HI participants are wearing HAs, so that there is no need to strongly change the SPL.

Authors	Speech	Noise
Arweiler and Buchholz [7]	n/a	60 dB SPL
Bradley *et al.* [25]	55 dB SPL	n/a
Culling and Mansell [54]	n/a	65 dB(A)
Drullman and Bronkhorst [61]	65 dB(A)	n/a
Duquesnoy and Plomp [62]	n/a	52.5 dB(A)
Ericson and McKinley [66]	73 dB SPL	n/a
Hawley *et al.* [96]	62 dB(A)	n/a

Table E.1 – Speech and noise levels used in some of the reported studies for NH subjects. Taken from [44].

Authors	Speech	Noise
Arweiler and Buchholz [7]	n/a	70 dB SPL
Drennan *et al.* [60]	70 dB(A)	70 dB(A)
Duquesnoy and Plomp [62]	n/a	52.5 dB(A)
Keidser *et al.* [114]	65 dB SPL	n/a
Köbler and Rosenhall [123]	73 dB SPL	n/a
Lewis *et al.* [136]	n/a	65 dB(A)
Picou *et al.* [187]	66 dB(A)	n/a
Simpson *et al.* [215]	60 dB(A)	n/a
Thibodeau [229]	84 dB(A)	n/a

Table E.2 – Speech and noise levels used in some of the reported studies for aided HI subjects. Taken from [44].

In the experiments related to speech intelligibility, some masking noises are played in the HAs of the listeners. First, one has to decide whether the sound is rendered diotically (exact same noise in both ears), decorellated (same nature of noise in both ears, but instantaneously different) or spatialized in certain azimuths. In recognition tests involving some loudspeakers, it is common to play the speech in the frontal loudspeaker, and use all the others together for the noise [7, 25, 114, 136, 229]. When intelligibility tests are conducted with headphones, several methods of rendering have been reported. Begault and Erbe [15] use a diotic babble noise with speech either presented diotically or spatialized at a variety of azimuths. Ericson and McKinley [66] play diotic and spatialized speech in 3 combinations of masking noise: diotic (correlation coefficient equal to 1), ambient (played through an array of loudspeaker, correlation coefficient of about 0.3) and combined diotic and ambient noise. They conclude that the speech understanding is the best in the case of an ambient noise and the worse for a diotic masker. Yousefian *et al.* [251] spatialize the speech signal at 0° and a babble noise at certain consecutive azimuths, while mentioning that a point source of babble noise is not a realistic rendering, but that it is extensively used in the literature.

E.1.3 Hardware

Hearing aids

In the reviewed studies involving HAs, 2 major scenarios can be distinguished: either the test is done with the usual HAs of the participants [123, 215, 229], or a new pair of HAs is provided, fitted accordingly to the NAL-NL1 recommendation [29] or the DSL prescription [213], and made beforehand available for several weeks for habituation. In fact, when the test does not use the original HAs of the listeners, there is the need for a certain time of accommodation (see e.g. [60, 114] for explanations). Many studies are actually working with one type of HAs only [60, 114, 136, 187], but those devices are specifically fitted for the test, with a long period of accommodation for all subjects.

When it comes to the specific signal processing features in the HAs, the strategy is varying between studies, depending on what is investigated, as summarized in Table E.3. For the details on these embedded features, as such as their potential effects on speech intelligibility, the reader should refer to Chapter 1.1.2 and 1.2.2.

Authors	Assessed effect(s)	Amplification	Directivity	Feedback cancellation	Noise reduction
Keidser *et al.* [114]	Low-frequency gain and venting	WDRC	ON	n/a	ON/OFF
Köbler and Rosenhall [123]	Bilateral and unilateral fittings	Linear and WDRC	OFF	n/a	n/a
Ibrahim *et al.* [104]	Binaural wireless technology	WDRC	OFF	OFF	OFF
Picou *et al.* [187]	Directional microphone	n/a	ON	ON	ON
Simpson *et al.* [215]	Audibility	Linear	n/a	OFF	OFF
Thibodeau [229]	Adaptive FM	WDRC	ON	ON	ON

Table E.3 – Activation and deactivation of signal processing features available in the HAs. Taken from [44].

The table confirms that there is no conventional rule governing the activation/deactivation of the signal processing features. Nevertheless, the adaptive signal processing is switched off in all studies, because it results in a variability among stimuli that one cannot control. Moreover, adaptive processing may introduce some left and right gain mismatches than can lead to some spatial cue distortions. The aided listeners do not keep their own earmolds in all the reported experiments. Thus, Simpson *et al.* [215] suggest to block the vent of the listeners' molds in order to minimize the sound leakage. Ibrahim *et al.* [104] provide new earmolds with no venting. The participants tested in the study of Keidser *et al.* [113] have earmolds with venting between 1 to 3 mm (certain subjects even have open earmolds).

DAI

The DAI allows to input an electrical signal directly in the HAs. This generally deactivates one microphone of the HA, because the input signal is treated as if it has been picked up by the microphone. The use of the DAI has been reported in 3 reviewed studies [37, 215, 229]. Simpson *et al.* [215] input the speech signal via the DAI of a unique model of HAs. They underline that the primary advantage of resorting to the DAI is that it avoids any feedback (microphones deactivated), and thus enables some higher gains, which would not be possible with the activated microphones.

The control of the SPL of the input signal can be managed with the suggested procedure by Chung *et al.* [37]. The HA is programmed with a linear amplification and its frequency response matches "*the gain targets recommended by the National Acoustic Laboratory fitting prescription in the 2cc coupler for the average hearing loss of the hearing-impaired listeners*". Then, a speech-shaped noise at 70 dB SPL is emitted by a loudspeaker, and the frequency response of the HA in the 2cc coupler is measured and stored. Finally, the same stimulus is input via the DAI, and the gain is adjusted until the frequency response of the HA in the 2cc coupler is the same as the one recorded with the loudspeaker. The authors conclude that "*this procedure ensured that the input to the hearing aid were equivalent to 70 dB SPL if the stimuli were presented acoustically to the microphone. It also verified that presenting test signals via DAI connections did not alter the frequency responses of the hearing aid compared to acoustic microphone inputs*". Of course, this procedure highly depends on the model of HAs, but not on the various internal fittings.

GUI

In a majority of study, no GUI is available for listeners. They are just asked to repeat what they have understood to an examiner present during the experiment. The only exceptions are in [54] that asks listeners to both write and tell what they have heard, in [60] that demands listeners to choose one sentence among 4, in [88] that asks them to circle the correct words among 10, in [96] that just asks them to write the sentence, and in [215] that makes them press the button corresponding to the perceived consonants on a screen among 16 possibilities.

E.1.4 Discussion

This literature review gives insight into the design of the intelligibility test. It is thought that 2 different experiments should be performed, as discussed in Table E.4.

These additional remarks can be formulated:

- Clinical trials on children are difficult to conduct. A balanced panel of young and old adults, with a congenital or a post-childhood HRL would be desired, so as to cover the major causes and effects of hearing impairment. Also, various degrees of HRLs should be considered. A control group of NH participants sounds interesting as well,

Experiment	Principle	Justification
FM-only	The stimuli are spatialized in one of the 5 spatial sectors and played through the HAs via the DAI. Some other stimuli are not spatialized, so as to represent the current rendering of the WMS.	This experiment allows to compare the intelligibility obtained with the current diotic rendering and the suggested spatialized rendering, in the FM-only mode, introduced in Chapter 1.4.1.
FM+M	The stimuli are simultaneously spatialized in one of the 5 considered locations and played in the corresponding loudspeaker, to be captured by the HA microphones. Some other stimuli are diotic, corresponding to the current rendering. Additionally, they are played by a random loudspeaker at the same time.	This experiment is similar to the first one, except that it allows to compare the effect of the FM+M rendering on the speech understanding, introduced in Chapter 1.4.1.

Table E.4 – The 2 intelligibility experiments of the proposed clinical trial. Taken from [45].

- The selection criteria suggested by Lewis *et al.* [136] and the conditions about symmetry should be retained, as long as they are not too restrictive to find a convenient number of participants,
- The tested SNRs cannot be the same among the various degrees of HRLs. Moreover, one has to avoid the ceiling and floor effects. Therefore, a method inspired by the one of Picou *et al.* is relevant, i.e. a set of different possible SNRs that are adapted to each subject,
- A loudness equalization between the diotic and spatialized stimuli is mandatory to prevent any bias,
- The use of a meaningful database with a large lexicon is the most pertinent approach for 2 reasons. First, it resembles the conditions of real life. Second, the cognitive processing is prominent for HI subjects, especially for severe and profound HRLs (see Appendix A.2.2). If a bias would occur, it would concern both modes or rendering,
- The resort to speech-shaped noise is appealing because it is very common and easy to generate. The main drawback is that it is not representative of what would be perceived in a classroom, which would be closer to a babble noise. On the other hand, the fluctuations inherently present in the babble noise makes it less stable and thus less reliable and reproducible,
- If diotic masking noise is used, it is obvious that spatialized speech stimuli will be better understood than diotic ones, as a consequence of the SRM (Appendix A.2.2). On the other hand, there is no reason to spatialize the noise in a specific direction. Therefore, it is thought that a decorellated noise is the most relevant type of maskers, since it is not perceived as a point source inside the head. Another possibility is to provide a noise spatialized in all sectors,
- The deactivation of the directivity and noise reduction algorithms is compulsory to avoid any risk of SNR improvements,

- The question of whether only Phonak HAs should be used is pertinent. There are numerous pros, such as the possibility to require a single fitting software, and the knowledge of all the technical specifications related to those HAs. Also, the risk of impedance mismatches between models is quite reduced. Finally, the statement of a comfortable level is made easier, as shown in Chapter 5.1.5. The main con is related to the consequent difficulty in finding a sufficient number of participants having the desired HAs,

- The DAI is used in the clinical trial, which means that all HA models must incorporate one,

- All of the reviewed studies including HAs are performed with HI users only. The question is whether it makes sense to ask NH subjects to wear HAs. This idea is supported by the need for having the same hardware for each participant, so that some relevant comparisons can be done. A defined setting must be fixed for all NH listeners. One also has to ensure that the perception of the direct sound is made as low as possible, e.g. by means of some occluded earmolds,

- No GUI should be made available for participants, especially because some elderly are included in the clinical trial, and they could encounter difficulties with it.

These conclusions are now completed with the requirements from the localization test.

E.2 Localization test

A review of 25 previously conducted localization tests, with NH and HI subjects, aided or not, and for real and virtual sound sources has been conducted. Certain test characteristics are the same as in the intelligibility tests, and are thus not detailed in what follows.

E.2.1 Subjects

Sample size

The reviewed studies involve an average number of 15 HI subjects (SD = 7), the minimum number being 7 [60], and the maximum number is 30 [218]. An average sample size of 10 NH subjects (SD = 7) is reported (a minimum of 4 in [20, 97, 142] and a maximum equal to 30 in [115]). Once again, the homogeneity of the results over NH participants allows to reduce the required sample size.

Age

It is likely that the influence of age is weaker for localization than for intelligibility, as the underlying cognitive processes involved in a localization task are of less importance.

Hearing and health

As for intelligibility test, symmetry of HRLs is required in almost all studies reporting a localization test, even for aided HI participants. The most common symmetry condition is less than 15 dB difference for all frequencies available on the audiogram [20, 60, 113, 114, 116, 180]. Ibrahim *et al.* [104] demand interaural variations lower than 10 dB at each frequency, while only 20 dB are asked by Köbler and Rosenhall [123], and by Whitmer *et al.* [244]. Only Lorenzi *et al.* [142] base their threshold on the average left and right audiograms.

E.2.2 Procedure

Task

Over the 25 reviewed studies, 2 major localization tasks and scenarios have been identified:

- In 13 studies, the listeners do not see any loudspeaker (hidden loudspeakers, blindfolded subjects, experiments in darkness, or no loudspeaker at all). They are asked to report the perceived incidence direction. The majority of the studies involving virtual sound sources (i.e. via headphones) employs this method [14, 17, 27, 31, 147, 243, 246],

- In 12 studies, the subjects are facing several visible loudspeakers and are asked to report which loudspeaker has originally emitted the sound. In almost all cases, the sound is played by the loudspeakers. There are 2 noticeable exceptions. In [233], the spatialized sound is played through headphones and subjects are asked to match the virtual source with a loudspeaker. In [37], binaural recordings of loudspeakers are played through headphones and the same task is conducted.

Subject feedback

Several means of collecting the answers of the participants have been reported:

- A simple verbal indication is the most used technique. In 16 studies over 25, an examiner is collecting the subjects' answers, be it a loudspeaker number, a perceived angle of incidence, or something else,

- A head tracking system is used in [20, 27, 31, 104]. In this case, the listener is asked to rotate the head toward the perceived direction, and the corresponding head orientation is recorded,

- Answers via a touch screen [14, 60],

- A magnetic search coil technique that measures the movements of the eyes is used in [97] and [244]. The subjects are asked to generate a quick eye movement in the perceived direction. The use of an electromagnetic stylus pointer is reported in [55] and in [147].

Head motion

The question of whether the listeners are allowed to turn the head during the stimulus presentation is related to the dynamic cues (Appendix A.2.1). In 15 of the reviewed studies, head motions are forbidden, i.e. listeners do not have access to these cues. Several techniques have been reported in the literature to this end:

- In the majority of studies [123, 147, 173, 233, 235, 243, 244, 246], the subjects are only asked to fix the 0° azimuth and not move their head, which is checked by the examiner,
- A head lamp worn by the listeners is reported in [113, 114, 115, 116]. In this case, the beam must focus on the 0° direction during all stimulus presentations,
- D'Angelo *et al.* [55] use a chin rest,
- Hofman *et al.* [97] resort to a head rest.

Training

A training session is included in a large majority of the reviewed studies. Its goal is to make listeners familiar with the experiment before starting the evaluation, so as to avoid any *learning effect*, i.e. an improvement of performance that occurs during the experiment. Various methods are reported for training:

- The most usual method is to train listeners through a certain number of trials at the beginning of the experiment, without giving any feedback on the accuracy of their answers [14, 17, 104, 113, 123, 142]. This is primarily achieved to check and ensure that subjects have well understood the task they have to perform,
- Bronkhorst [27] and Majdak *et al.* [147] also asks listeners to practice a certain number of trials, but they give a feedback after each answer to inform the subjects of the amount of error they just did, or to show them the actual location of the sound source,
- In [37], subjects are played some binaural recordings in each available location, and the examiner indicates the corresponding physical location. No trial is done before starting the experiment,
- Certain authors [173, 235] combine the 2 last reported methods, first presenting all the possible locations, then making listeners practice some trials. They finally give a feedback after each trial.

Whereas it makes sense to present all the available locations and provide feedbacks for real sound sources, this method is questionable when it comes to virtual stimuli. Indeed, the perceived spatial position associated with a virtual sound source is something purely subjective, which depends on several factors, such as the filter quality, the HRTFs of each tested subject,

their hearing characteristics... One might not "force" listeners to match a stimulus filtered with a certain pair of spatial filters with a given physical location.

Scoring localization performance

The test protocol and the way subjects' localization performance is scored are intrinsically related. That is the reason why one should think of the desired assessment method while designing the test. In the literature, the main scores are the following:

- The mean absolute error, which is defined as the average absolute value of the difference between the real and perceived incidence directions in degrees [14, 20, 27, 37, 55, 147, 218, 233, 243, 246],
- The RMS error, which is close to the mean absolute error except that it gives more weight to strong localization errors [60, 113, 114, 115, 116, 142, 173, 235],
- The mean signed error, which is usually associated with the 2 previous errors to check whether there is a significant bias to either side of the head [113, 114, 115, 116],
- The score given as a percentage: the proportion of the correctly identified loudspeakers in different listening conditions [123], or the proportion of the correct answers in each direction [180, 187].

E.2.3 Stimuli

Nature

A wide variety of test signals is used for localization experiments. The 2 main families of signal are noise and speech. Over the 25 reviewed studies, 15 resort to various kinds of noise stimuli (broadband white or pink noise, band-limited noise, speech-shape noise, pulsed noise...) and 10 use a speech signal.

SPL

Table E.5 summarizes the SPL used in various localization experiments, for NH and HI listeners. It can be noticed that the most common level delivered to NH listeners is 70 dB SPL (around 65 dB(A)), while it is lower for HI listeners, mainly 65 dB SPL. This can look odd, but one has to keep in mind that HI listeners are usually tested with HA. A lower emission level is probably chosen for HI listeners, so that the dynamic compression is limited.

Roving

The *roving* is a random and small variation of the static level across stimuli and locations. It is implemented in order to minimize the risk that participants rely on some perceived differences of loudness among locations to infer the position of the sound source [37, 113, 142]. This can be due to calibration biases, loudness difference between the spatial filters used, the HRL and

Authors	NH listeners	HI listeners
Begault *et al.* [14]	60 dB(A) (speech)	-
Best *et al.* [20]	55 dB SPL	65 dB SPL
Begault and Wenzel [17]	70 dB SPL	-
Bronkhorst [27]	60 dB(A)	-
Chung *et al.* [37]	70 dB SPL	-
D'Angelo *et al.* [55]	70 dB SPL	-
Drennan *et al.* [60]	-	70 dB(A) (speech)
Hofman *et al.* [97]	60 dB SPL	-
Ibrahim *et al.* [104]	-	76 dB SPL
Keidser *et al.* [113, 114, 116]	-	65 dB SPL
Köbler and Rosenhall [123]	-	73 dB SPL
Lorenzi *et al.* [142]	-	70 dB SPL
Mueller *et al.* [173]	63 dB SPL	-
Noble *et al.* [180]	-	50 dB SPL
Picou *et al.* [187]	-	66 dB(A) (speech)
Sockalingam *et al.* [218]	60 dB SPL	-
Van den Bogaert *et al.* [233, 235]	65 dB(A) (speech)	-
Wenzel *et al.* [243]	70 dB SPL	-
Whitmer *et al.* [244]	55 dB(A)	55 dB(A)
Wightman and Kistler [246]	70 dB SPL	-

Table E.5 – SPL of stimuli used in various localization test. Taken from [45].

the HA fittings of the participants... A wide range of roving is applied in the different reviewed studies: 2 dB in [142, 180], 5 dB in [147], and 6 dB in [116, 233, 235]. Note that a roving of e.g. 6 dB usually means that the stimuli are played at any random value between -3 dB and +3 dB relative to the static SPL.

E.2.4 Hearing aids

As mentioned in Chapter 1.2.1, certain signal processing features available in HAs are known to be detrimental to sound localization. Like in the intelligibility tests, the prior deactivation of some algorithms depends on what is investigated. Table E.6 present a summary of the fittings adopted in the literature. Overall, it appears that the WDRC and feedback cancellation are often kept active, while the other algorithms tend to be discarded.

Regarding the molds and vents, there is no specific rule reported in the literature. Open earmolds are used in [235], while closed earmolds with no vent are employed in [104]. Vents between 1 to 3 mm are reported in [20, 114, 115, 116]. Chung *et al.* [37] suggest to use compressible foam tips that are coupled to the HAs, in order to replace the earmolds.

Surprisingly, studies reporting localization tests with HAs are somehow less constraining in terms of HA experience. In [20], only 4 subjects over 11 are experienced with HAs, but an

Authors	Assessed effect(s)	Amplification	Directivity	Feedback cancellation	Noise reduction
Best *et al.* [20]	CIC and BTE HAs	WDRC	OFF	n/a	OFF
Chung *et al.* [37]	Directional microphone	Linear	ON	n/a	n/a
Drennan *et al.* [60]	Phase-preserving amplification	Linear	n/a	n/a	n/a
Ibrahim *et al.* [104]	Binaural wireless technology	WDRC	OFF	OFF	OFF
Keidser *et al.* [114]	Low-frequency gain and venting effects	WDRC	ON/OFF	n/a	ON/OFF
Keidser *et al.* [115]	Directional microphone	WDRC	OFF	ON	OFF
Keidser *et al.* [116]	Interaural gain mismatch	WDRC	OFF	n/a	OFF
Köbler and Rosenhall [123]	Bilateral and unilateral fittings	WDRC/Linear	OFF	n/a	n/a
Picou *et al.* [187]	Directional microphone	n/a	ON	ON	ON
Van den Bogaert *et al.* [235]	BTE, ITE and ITC HAs	n/a	n/a	ON	OFF

Table E.6 – Activation and deactivation of signal processing features in the HAs. Taken from [45].

accommodation period of 4 to 6 weeks is scheduled before beginning the test. In [218], 16 of the 30 listeners are experienced, and an accommodation period of 2 weeks is included. 21 listeners over 23 are experienced in [114], with an acclimatization duration of 4 weeks, and 16 over 21 listeners are experienced in [115], with an acclimatization period of 3 weeks. For the 9 other studies involving bilateral HAs, all listeners are experienced for a long time (from several months to several years).

E.2.5 Discussion

From this literature review, several ideas arise. First, it is thought that 4 different experiments should be performed, as detailed in Table E.7.

Some additional remarks can be made:

- Involving children in this study would help know if the spatial filters based on an adult head is adapted to young listeners,
- A mix of real and virtual sound sources is considered for the clinical trial. A common task in every experiment is mandatory to allow some easy comparisons. Asking listeners, particularly unexperienced and/or HI subjects, to match a virtual stimulus location with a physical position seems to be a tough task, especially without a long training period. Also, one has to wonder whether the subjects can give any perceived direction, or if they have to choose between a finite set of possible sound source locations,

Experiment	Principle	Justification
Unaided	Listeners are unaided. The stimuli are played in one of the 5 loudspeakers located in the same physical positions as the spatial sectors of the algorithm.	This experiment gives insight into the basic localization performance of the participants.
Aided	Listeners are aided, with their usual HAs and fittings. The stimuli are played in one of the 5 loudspeakers located in the same physical positions as the spatial sectors.	This experiment serves to compare the effect of the day-to-day HA processing on the localization capabilities.
FM-only	The stimuli are spatialized in one of the 5 spatial sectors and played through the HAs via the DAI.	This experiment allows to investigate both the perception of binaural spatialization on HI subjects, and the possibility to render an adequate spatialized sound with HAs. Comparing the results with the ones of the second experiment, one can check whether the usual localization performance of the participants are preserved by the processing.
FM+M	The stimuli are simultaneously spatialized in one of the 5 considered locations and played in the corresponding loudspeaker, to be captured by the HA microphones.	This experiment evaluates the performance of subjects in the FM+M mode, introduced in Chapter 1.4.1. Comparisons with the third experiment allow to analyse which of the FM-only or FM+M mode is the most suitable for the spatialization rendering. With the results of the second experiment, one can also compare the contribution of the FM to the usual listening with the HA microphones.

Table E.7 – The different localization experiments of the clinical trial. Taken from [45].

- In order to facilitate the procedure for the participants, a verbal indication of the identified DOA is favored,
- For both intelligibility and localization tests, it is compulsory that the subjects do not move their head during the presentation of the real sound sources, because the spatialization is not dynamically processed. Although it is the most used and easiest technique, a simple order from the examiner does not seem to be sufficient, because head movements can be slight and not detected. Regarding the coarse spatial resolution considered in this study, the resort to some high-precision systems, such as a head tracking, a magnetic search coil or an electromagnetic pointer is not appropriate. The resort to a chin rest is thought to be the most efficient method,
- A training with several runs would help the listener before starting the test. No feedback should be given, because there is no true or false answers when one deals with binaural spatialization,
- The mean absolute error and RMS error are not suitable, since the listeners' answers should be simply reported in terms of spatial sector. A percentage of correct localization in each of the 5 spatial locations would be more convenient, but it gives no indication about the error size. Therefore it should be combined with an error computed on some numbers given for each sector (e.g. the targeted speech is located in the central sector

(numbered “0”), the perceived sector is the extreme left (numbered “2”), so the resulting error is 2-0 = 2 sectors. Similarly, one could define a signed mean error in this way,

- Speech is the typical signal that is processed by WMS. Thus, it is considered as the most suitable stimulus for the test. It would be better that the content of the speech could not be understood by listeners, to avoid them taking care to the meaning of the sentence, instead of concentrating on its spatial position. In order to ensure this, it is suggested to employ the SUS database,

- Roving should be included to reduce the risk of loudness-based localization,

- Although WDRC is known to distort the spatial cues (Chapter 1.2.1), the non-linear amplification can be considered as the most important processing of the HAs and should not be deactivated. Moreover, switching it off does not match a daily-use of HAs. This also supports the preservation of the participants’ earmolds. On the other side, directivity is not desired, because it is not available when the FM is working.

The following now reviews the preference-rating tests.

E.3 Preference rating

The content of this part is based on a review of 10 previously conducted subjective evaluations, with NH and HI subjects, aided or not. The topics related to subjects, stimuli and hardware do not change compared to part E.1 and E.2, and are thus not mentioned in the following.

E.3.1 Procedure

Standards

2 major methods for conducting some subjective evaluations have been noticed in the literature. The first consists in presenting successive audio stimuli to the listeners and asked them to grade them one after the other. With this technique, the subjects have to assess each stimulus in a purely absolute way, i.e. without comparing with the previous ones. In [14], the listeners are asked to rate the “perceived realism” of consecutive sound samples on a scale from 0 to 4. The authors notice no significant effect of the tested stimuli on the ratings of the listeners. They thus wonder whether the subjects had a common understanding of what “realism” meant. In [233], the participants must assess the perceived width of the stimuli among 4 possibilities. Again, there is no comparison between stimuli, which are rated one after the other. Sockalingam *et al.* [218] ask some HI subjects to rate the naturalness of the presented stimuli through HAs, using a rank from 1 to 6. The method of Drennan *et al.* [60] is a bit different in the sense that it is not based on some presented stimuli, rather on the experience of HI subjects wearing HAs in their daily life. After 16 weeks of living with a

particular HA fitting, the subjects are asked to answer 49 questions by giving a mark between 1 and 10.

It is often difficult to analyze the results of a numerical evaluation, as each subject has their own scale. For example, if a scale between 0 and 10 is proposed, some listeners would only exploit the 4-6 range, while some others would give marks in the whole range. Furthermore, the consistency of the ratings along the test can be poor. Indeed, it is common that a certain learning effect occurs, leading to some significant changes in the given grades from the beginning to the end of the test.

In the second method, the listeners are played several stimuli (2 or more). The task to be performed then varies depending on the procedure: the subjects can be asked to indicate their preference, e.g. by classifying the stimuli according to some specific criteria, or by grading them. This is achieved in [187], where the examiner switches between 2 types of processing, and the listeners are free to listen to both stimuli as much as they need. Then, the examiner asks guided questions to the subjects, but the answers are not collected (i.e. the questions are only used to make listeners attentive to some specific attributes in the sound). Eventually, the listeners have to mention which of the 2 stimuli is preferred, according to the attributes they just focused on. This procedure is called the *paired comparison* [175]. Nielsen [175] indicates that for elderly and/or untrained persons, a paired comparison should provide more reliable results than the first absolute method. Hence, this procedure is often judged as more sensitive to detect significant differences between certain processing.

2 other experiments are reported in [141]. In the first one, a paired comparison is performed, and the listeners have to switch between 2 different echo-cancellation processing. They have to report which of the 2 stimuli provides the best realism, and then have to rate the amount of differences they perceive within a scale from 1 to 5. This last technique is also well established and called *similarity rating* [175]. In the second experiment, the participants have to switch between 6 different stimuli, corresponding to various spatialization processing. They are asked to order the 6 cases regarding different attributes (intelligibility, immersion...). In Le Bagousse *et al.* [133], 8 different music encodings are played to the subjects, who must grade them simultaneously according to 3 sound attributes. A reference (i.e. the original encoding), and 3 anchors (i.e. the worst conditions), one for each attribute, are hidden among the 8 stimuli. Note that this approach resembles the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) methodology that is recommended by the International Telecommunication Union for the assessment of audio quality (recommendation BS. 1534-1 [107]).

Sound attributes

Bech [13] defines an *auditory attribute* as "*a perceptual characteristic of a sound stimulus*". As reported in [132], there is no standard sound attributes that are used in all studies. On the contrary, each author usually employs their own terms and definitions. Colomes *et al.* [38] have gone through an extensive review of the literature dealing with the evaluation of quality, and propose a classification of the redundant attributes in 3 categories:

- Defaults: Attributes for the evaluation of "*interfering elements of nuisances present in a sound*", e.g. noise, distortion, hum, hiss, clipping, disruption...,
- Space: Attributes related to the spatial perception: depth, width, localization, spatial distribution, reverberation, spatialization, distance, envelope, immersion...,
- Timbre: Attributes for the assessment of the temporal and frequency content: brightness, coloration, clarity, hardness, richness, sharpness, realism, naturalness, pleasantness...

It is possible to associate each chosen attribute with a scale going from a minimum to a maximum mark [14, 218, 233]. One should ensure that the definition of each attribute is clear and well understood by all subjects.

Instead of asking subjects to evaluate the attributes by some simple words, some researchers resort to more general questions that are supposed to be easier for the listeners. Picou *et al.* [187] ask guided questions before the subjects choose which of the 2 presented stimuli they prefer. These questions are related to intelligibility ("*How well can you understand speech?*"), and spatial perception ("*How well can you tell where the sounds are coming from?*").

It is prominent to state that these evaluations do not interfere with the objective intelligibility and localization tests (Part E.1 and E.2). Indeed, a listener can succeed in correctly reporting the content and location of 2 speech stimuli that do not demand the same amount of effort. This can be revealed by such an investigation, where participants can give a better preference to the stimuli that have been easier to understand or localize [175]. Thus, after some objective assessments, Lopez *et al.* [141] ask the subjects to answer 4 questions in order to classify the 6 stimuli they had to compare. These questions are related to:

- Intelligibility: "*Do you understand the conversation easily? How much?*",
- Distance: "*Do you have a sense of distance when you move the speakers? How much?*",
- Immersion: "*Do you feel immersed in a real environment or room? How much?*",
- Overall preference: "*How much do you like each option? Order them according to your personal preference*".

One also has to mention the speech, spatial and qualities (SSQ) of hearing scale, written by Gatehouse and Noble [77]. It has been established to assess the disability encountered by HI subjects and the impact of several real-life factors. It consists in 49 questions, split in 3 sections: speech-hearing, spatial-hearing, and quality of hearing. The answer to each question is given by a score between 0 and 10. Some examples of interesting questions are reported here:

- Speech-hearing, question 13: "*You are with a group and the conversation switches from one person to another. Can you easily follow the conversation without missing the start of what each new speaker is saying?*",

- Spatial-hearing, question 2: "*You are sitting around a table or at a meeting with several people. You can't see everyone. Can you tell where any person is as soon as they start speaking?*",
- Spatial-hearing, question 3: "*You are sitting in between two people. One of them starts to speak. Can you tell right away whether it is the person on your left or your right, without having to look?*"
- Spatial-hearing, question 14: "*Do the sounds of things you are able to hear seem to be inside your head rather than out there in the world?*",
- Hearing qualities, question 18: "*Do you have to put a lot of effort to hear what is being said in conversation with others?*".

Training

No specific training sessions are mentioned in the reviewed studies, except in [175]. Generally, the acquisition of the subjects' answers starts immediately when the examiner is sure that the task is well understood by the listeners. In [14], no information is given about the attributes to evaluate, and this probably explains the absence of significant effect in the results. That is the reason why Nielsen [175] underlines the importance of having clear and short instructions of the tasks to be performed. Questions and comments should always be heard and answered. Nielsen includes a "warm-up" of 16 stimuli that are disregarded and followed by 4 runs of 16 stimuli for the evaluation.

E.3.2 Discussion

Subjective evaluations are known to be tough because of the wide variability that may occur between subjects. The reported literature helps guide the way this test should be done. In particular, the following remarks can be made:

- The rendering of the stimuli is planed to be done via some movies to bring more realism. This requires that the participants have no history of epilepsy or some other specific bad reactions related to the proximity to a video screen. Additionally, they should not suffer from a severe visual impairment. A vision screening test is not intended, but the subjects must be able to guarantee a correct vision when watching a screen, with the wearing of glasses or not,
- In the clinical trial, the goal is to compare 2 different renderings (diotic or spatialized). Therefore, it is thought that a paired comparison is the most adapted approach,
- The use of numerical ratings (e.g. ask listeners to give a grade) should be avoided. A simple answer like an information about which rendering is preferred seems more straightforward and reliable,

- Concerning the sound attributes, it is suggested to resort to questions, rather than simple words with associated definitions. The preference over several and specific successive attributes should be asked. The attributes that are thought prominent for the current assessment are the following: clarity/intelligibility/ease of understanding, immersion, naturalness/reality/externalization, and an overall preference.

Bibliography

[1] S. M. Abel, C. Giguère, A. Consoli, and B. C. Papsin. The effect of aging on horizontal plane sound localization. *J. Acoust. Soc. Am.*, 108(2):743–752, 2000.

[2] H. B. Abrams and J. Kihm. An introduction to marketrak IX: a new baseline for the hearing aid market, 2015. URL http://www.hearingreview.com/2015/05/introduction-marketrak-ix-new-baseline-hearing-aid-market/.

[3] T. Ajdler, C. Faller, L. Sbaiz, and M. Vetterli. Interpolation of head related transfer functions considering acoustics. In *Proceedings of the 118th AES Convention*. Audio Engineering Society, 2005.

[4] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *Proceedings of the workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 99–102. Institute of Electrical and Electronics Engineers, 2001.

[5] ANSI. S3. 22–2003 Specification of hearing aid characteristics. *American National Standards Institute*, 2003.

[6] J. Appleton and G. König. Improvement in speech intelligibility and subjective benefit with binaural beamformer technology, 2014. URL http://www.hearingreview.com/2014/10/improvement-speech-intelligibility-subjective-benefit-binaural-beamformer-technology/. 2016-03-08.

[7] I. Arweiler and J. M. Buchholz. The influence of spectral characteristics of early reflections on speech intelligibility. *J. Acoust. Soc. Am.*, 130(2):996–1005, 2011.

[8] World Medical Association. Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects, 1964.

[9] A. Awad, T. Frunzke, and F. Dressler. Adaptive distance estimation and localization in WSN using RSSI measures. In *Proceedings of the 10th Euromicro Conference on Digital System Design Architectures, Methods and Tools (DSD)*, pages 471–478. Institute of Electrical and Electronics Engineers, 2007.

[10] W. Balande. D4.2&3 Software implementation on Atmal ARM9 - Restricted. KTI deliverable, Phonak Communications AG, 2014.

[11] W. Balande. Localization & Spatialization Profiling - Restricted. Internal report, Phonak Communications AG, 2014.

[12] P. Barsocchi, S. Lenzi, S. Chessa, and G. Giunta. A novel approach to indoor RSSI localization by automatic calibration of the wireless propagation model. In *Proceedings of the 69th Vehicular Technology Conference (VTC)*, pages 1–5. Institute of Electrical and Electronics Engineers, 2009.

[13] S. Bech. Methods for subjective evaluation of spatial characteristics of sound. In *Proceedings of the 16th AES Conference*. Audio Engineering Society, 1999.

[14] D. Begault, E. Wenzel, A. Lee, and M. Anderson. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. In *Proceedings of the 108th AES Convention*, volume 5134. Audio Engineering Society, 2001.

[15] D. R. Begault and T. Erbe. Multichannel spatial auditory display for speech communications. *J. Audio. Eng. Soc.*, 42(10):819–826, 1994.

[16] D. R. Begault and E. M. Wenzel. Techniques and applications for binaural sound manipulation. *Int. J. Aviat. Psychol.*, 2(1):1–22, 1992.

[17] D. R. Begault and E. M. Wenzel. Headphone localization of speech. *Hum. Factors*, 35(2): 361–376, 1993.

[18] C. Benoît, M. Grice, and V. Hazan. The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Commun.*, 18(4):381–392, 1996.

[19] R. Bentler and Li-K. Chiou. Digital noise reduction: an overview. *Trends Amplif.*, 10(2): 67–82, 2006.

[20] V. Best, S. Kalluri, S. McLachlan, S. Valentine, B. Edwards, and S. Carlile. A comparison of CIC and BTE hearing aids for three-dimensional localization of speech. *Int. J. Audiol.*, 49(10):723–732, 2010.

[21] V. Best, C.R. Mason, G. Kidd Jr, N. Iyer, and D.S. Brungart. Better-ear glimpsing in hearing-impaired listeners. *J. Acoust. Soc. Am.*, 137(2):EL213–EL219, 2015.

[22] J. Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[23] A. Bohnert, M. Nyffeler, and A. Keilmann. Advantages of a non-linear frequency compression algorithm in noise. *Eur. Arch. Otorhinolaryngol.*, 267(7):1045–1053, 2010.

[24] A. W. Boyd, W. M. Whitmer, J. J. Soraghan, and M. A. Akeroyd. Auditory externalization in hearing-impaired listeners: The effect of pinna cues and number of talkers. *J. Acoust. Soc. Am.*, 131(3):EL268–EL274, 2012.

[25] J. S. Bradley, R. D. Reich, and S. G. Norcross. On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility. *J. Acoust. Soc. Am.*, 106(4):1820–1828, 1999.

[26] M. S. Brandstein. Time-delay estimation of reverberated speech exploiting harmonic structure. *J. Acoust. Soc. Am.*, 105(5):2914–2919, 1999.

[27] A. W. Bronkhorst. Localization of real and virtual sound sources. *J. Acoust. Soc. Am.*, 98 (5):2542–2553, 1995.

[28] D. Byrne and W. Noble. Optimizing sound localization with hearing aids. *Trends Amplif.*, 3(2):51–73, 1998.

[29] D. Byrne, H. Dillon, T. Ching, R. Katsch, and G. Keidser. NAL-NL1 procedure for fitting nonlinear hearing aids: characteristics and comparisons with other procedures. *J. Am. Acad. Audiol.*, 12(1):37–51, 2001.

[30] R. Carhart. Monaural and binaural discrimination against competing sentences. *J. Acoust. Soc. Am.*, 37(6):1205–1205, 1965.

[31] S. Carlile, C. Jin, and V. Van Raad. Continuous virtual auditory space using HRTF interpolation: Acoustic and psychophysical errors. In *Proceedings of the Pacific-Rim Conference on Multimedia (PCM)*, pages 220–223. Institute of Electrical and Electronics Engineers, 2006.

[32] B. Carty and V. Lazzarini. Frequency-domain interpolation of empirical HRTF data. In *Proceedings of the 126th AES Convention*. Audio Engineering Society, 2009.

[33] WHO Media centre. Deafness and hearing loss: Fact sheet N°300, 2015. URL http://www.who.int/mediacentre/factsheets/fs300/en/. 2016-07-05.

[34] CER-VD. Commission cantonale (VD) d'éthique de la recherche sur l'être humain, 2016. URL http://www.cer-vd.ch/. 2016-07-05.

[35] C. I. Cheng and G. H. Wakefield. Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space. In *Proceedings of the 107th AES Convention*. Audio Engineering Society, 1999.

[36] L. Cheng, C. D. Wu, and Y. Z. Zhang. Indoor robot localization based on wireless sensor networks. *IEEE Trans. Consum. Electron.*, 57(3):1099–1104, 2011.

[37] K. Chung, A. C. Neuman, and M. Higgins. Effects of in-the-ear microphone directionality on sound direction identification. *J. Acoust. Soc. Am.*, 123(4):2264–2275, 2008.

[38] C. Colomes, S. Le Bagousse, and M. Paquier. Families of sound attributes for assessment of spatial audio. In *Proceedings of the 129th AES Convention*. Audio Engineering Society, 2010.

[39] G. Courtois. D1.2 Psychoacoustic requirements - Restricted. KTI deliverable, 2012.

[40] G. Courtois. D2.1 Binaural sound source localization processing - Restricted. KTI deliverable, EPFL / Phonak Communications AG, 2013.

[41] G. Courtois. D3.1 Binaural restitution with spatial cues relative to the listeners - Restricted. KTI deliverable, EPFL, 2014.

[42] G. Courtois. D3.2 Tests in-lab and in-vivo for the BHA(S) design - Restricted. KTI deliverable, EPFL, 2014.

[43] G. Courtois. Tuning of the localization parameters - Restricted. Internal report, EPFL, 2014.

[44] G. Courtois. Design of an intelligibility test for the evaluation of the BHA(S) processing - Restricted. Internal report, EPFL, 2015.

[45] G. Courtois. Design of a sound localization test for the evaluation of the BHA(S) processing - Restricted. Internal report, EPFL, 2015.

[46] G. Courtois. Evaluation of a binaural spatialization method for hearing aids, in terms of speech intelligibility, speaker localization and subjective preference: test protocol - Restricted. Technical report, EPFL / Phonak Communications AG, 2016.

[47] G. Courtois and Sonova AG. ClinicalTrial.gov - US National Institutes of Health, 2016. URL https://clinicaltrials.gov/ct2/show/NCT02693704.

[48] G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, W. Balande, and Phonak AG. Spatial hearing - Audio localization and spatialization (provisional title) - pending. EP 2015 051265.

[49] G. Courtois, P. Marmaroli, Y. Oesch, and W. Balande. Spatial Hearing - Audio localization - Restricted. Internal invention report E13054, EPFL / Phonak Communications AG, 2013.

[50] G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande. Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions. In *Proceedings of the 136th AES Convention*. Audio Engineering Society, 2014.

[51] G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande. Binaural hearing aids with wireless microphone systems including speaker localization and spatialization. In *Proceedings of the 138th AES Convention*. Audio Engineering Society, 2015.

[52] G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande. Development and assessment of a localization algorithm implemented in binaural hearing aids. In *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*. European Association for Signal Processing, 2015.

[53] C. J. Crandell and J. J. Smaldino. Improving classroom acoustics: utilizing hearing-assistive technology and communication strategies in the educational setting. *Volta Review*, 101(5):47–62, 1999.

[54] J. F. Culling and E. R. Mansell. Speech intelligibility among modulated and spatially distributed noise sources. *J. Acoust. Soc. Am.*, 133(4):2254–2261, 2013.

[55] W. R. D'Angelo, R. S. Bolia, P. J. Mishler, and L. J. Morris. Effects of CIC hearing aids on auditory localization by listeners with normal hearing. *J. Speech Lang. Hear. R.*, 44(6): 1209–1214, 2001.

[56] M. S. Datum, F. Palmieri, and A. Moiseff. An artificial neural network for sound localization using binaural cues. *J. Acoust. Soc. Am.*, 100(1):372–383, 1996.

[57] Collège National d'Audioprothèse. Hearing In Noise Test (HINT) - French database, 2006.

[58] H. Dillon. *Hearing Aids*. Thieme, 2012.

[59] P. L. Divenyi, P. B. Stark, and K. M. Haupt. Decline of speech understanding and auditory thresholds in the elderly. *J. Acoust. Soc. Am.*, 118(2):1089–1100, 2005.

[60] W. R. Drennan, S. Gatehouse, P. Howell, D. Van Tasell, and S. Lund. Localization and speech-identification ability of hearing-impaired listeners using phase-preserving amplification. *Ear and Hear.*, 26(5):461–472, 2005.

[61] R. Drullman and A. W. Bronkhorst. Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *J. Acoust. Soc. Am.*, 107(4):2224–2235, 2000.

[62] A. J. Duquesnoy and R. Plomp. Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis. *J. Acoust. Soc. Am.*, 68(2):537–544, 1980.

[63] B. Edwards. The future of hearing aid technology. *Trends Amplif.*, 11(1):31–46, 2007.

[64] T. M. Elliott and F. E. Theunissen. The modulation transfer function for speech intelligibility. *PLoS. Comput. Biol.*, 5(3), 2009.

[65] Embracehearing. A 25-year rise in Hearing aid prices- & the embrace solution, 2012. URL http://www.embracehearing.com/blogs/hearingaidsnews/5687052-a-25-year-rise-in-hearing-aid-prices-the-embrace-solution. 2016-07-05.

[66] M. A. Ericson and R. L. McKinley. The intelligibility of multiple talkers separated spatially in noise. Technical report, DTIC Document, 2001.

[67] J. Escolano, N. Xiang, J. M. Perez-Lorenzo, M. Cobos, and J. J. Lopez. A Bayesian direction-of-arrival model for an undetermined number of sources using a two-microphone array. *J. Acoust. Soc. Am.*, 135(2):742–753, 2014.

[68] D. A. Fabry. Noise reduction with FM systems in FM/EM mode. *Ear and Hear.*, 15(1): 82–86, 1994.

[69] C. Faller. *Signal processing for audio and acoustics*. EPFL, 2011.

[70] X. Falourd, P. Marmaroli, F. Marquis, and Y. Oesch. BHA(L&S) submission form to the Kommission für Technologie und Innovation (KTI), 2012.

[71] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen. Informed direction of arrival estimation using a spherical-head model for hearing aid applications. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing ({ICASSP}) 2016*, pages 360–364. Institute of Electrical and Electronics Engineers, 2016.

[72] J. M. Festen and R. Plomp. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am.*, 88(4): 1725–1736, 1990.

[73] H. Fischer. Sources localization in wireless microphone system for people with hearing loss - Restricted. Master thesis, University of Applied Sciences Western Switzerland, 2011.

[74] F. P. Freeland, L. W.P. Biscainho, and P. S. R. Diniz. Interpositional transfer function for 3d-sound generation. *J. Audio. Eng. Soc.*, 52(9):915–930, 2004.

[75] F. Fujii, N. Hogaki, and Y. Watanabe. A simple and robust binaural sound source localization system using interaural time difference as a cue. In *Proceeding of the International Conference on Mechatronics and Automation (ICMA) 2016*, pages 1095–1101. Institute of Electrical and Electronics Engineers, 2013.

[76] M. B. Gardner. Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space. *J. Acoust. Soc. Am.*, 45(1):47–53, 1969.

[77] S. Gatehouse and W. Noble. The speech, spatial and qualities of hearing scale (SSQ). *Int. J. Audiol.*, 43(2):85–99, 2004.

[78] S. A. Gelfand. *Hearing: An introduction to psychological and physiological acoustics*. Marcel Dekker Inc, 1998.

[79] Genesis. GENESIS - loudness online, 2009. URL http://genesis-acoustics.com/en/loudness_online-32.html. 2016-06-17.

[80] E. J. L George, J. M. Festen, and T. Houtgast. Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 120(4):2295–2311, 2006.

[81] E. L. J. George, J. M. Festen, and T. Houtgast. The combined effects of reverberation and nonstationary noise on sentence intelligibility. *J. Acoust. Soc. Am.*, 124(2):1269–1277, 2008.

[82] B. R. Glasberg and B. C. J. Moore. A model of loudness applicable to time-varying sounds. *J. Audio. Eng. Soc.*, 50(5):331–342, 2002.

[83] H. Glyde, S. Cameron, H. Dillon, L. Hickson, and M. Seeto. The effects of hearing impairment and aging on spatial processing. *Ear and Hear.*, 34(1):15–28, 2013.

[84] J. S. Gravel, N. Fausel, C. Liskow, and F. Chobot. Children's speech recognition in noise using omni-directional and dual-microphone hearing aid technology. *Ear and Hear.*, 20 (1):1–11, 1999.

[85] S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay. *Speech processing in the auditory system.* Springer, 2004.

[86] R. Grewal and J. Irwin. Innovative ways hearing aids can be improved for clinical use: a literature review. *Scott. Uni. Med. J.*, 1(7), 2012.

[87] M. Guo, S. H. Jensen, and J. Jensen. Evaluation of state-of-the-art acoustic feedback cancellation systems for hearing aids. *J. Audio. Eng. Soc.*, 61(3):125–137, 2013.

[88] D. C Halling and L. E. Humes. Factors affecting the recognition of reverberant speech by elderly listeners. *J. Speech Lang. Hear. R.*, 43(2):414–431, 2000.

[89] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass. Signal processing in high-end hearing aids: state of the art, challenges, and future trends. *J. Appl. Sig. P.*, 2005:2915–2929, 2005.

[90] P. S. K. Hansen and Oticon A/S. Listening system adapted for real-time communication providing spatial information in an audio stream, 2013. EP 2 584 794 A1.

[91] A. Härmä and J. Huopaniemi. wfilter.c, 1997. URL http://legacy.spa.aalto.fi/software/warp/WarpTB/source/wfilter.c.

[92] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi. Frequency-warped signal processing for audio applications. *J. Audio. Eng. Soc.*, 48(11): 1011–1031, 2000.

[93] W. M. Hartmann and A. Wittenberg. On the externalization of sound images. *J. Acoust. Soc. Am.*, 99(6):3678–3688, 1996.

[94] K Hartung and A Raab. Efficient modelling of Head-Related Transfer Function. In *Proceeding of Acta Acust. united Ac.*, volume 82, page 88. Acta Acustica united with Acustica, 1996.

[95] D.B. Hawkins. Comparisons of speech recognition in noise by mildly-to-moderately hearing-impaired children using hearing aids and FM systems. *J. Speech Hear. Disord.*, 49(4):409–418, 1984.

[96] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn. Speech intelligibility and localization in a multi-source environment. *J. Acoust. Soc. Am.*, 105(6):3436–3448, 1999.

[97] P. M. Hofman, J. G. A. Van Riswick, and A. J. Van Opstal. Relearning sound localization with new ears. *Nat. Neurosci.*, 1(5):417–421, 1998.

[98] K. Hopkins and B. C. J. Moore. The effects of age and cochlear hearing loss on temporal fine structure sensitivity, frequency selectivity, and speech reception in noise. *J. Acoust. Soc. Am.*, 130(1):334–349, 2011.

[99] J. Huang, N. Ohnishi, and N. Sugie. Sound localization in reverberant environment based on the model of the precedence effect. *IEEE Trans. Instrum. Meas.*, 46(4):842–846, 1997.

[100] J. Huch and Ed. HHTM. Prevalence and future cost of hearing loss, 2016. URL http://hearinghealthmatters.org/theaudiologycondition/2016/prevalencefuturecosthearingloss/. 2016-03-08.

[101] J. Huopaniemi and M. Karjalainen. Review of digital filter design and implementation methods for 3-D sound. In *Proceedings of the 102nd AES Convention*. Audio Engineering Society, 1997.

[102] J. Huopaniemi, N. Zacharov, and M. Karjalainen. Objective and subjective evaluation of head-related transfer function filter design. *J. Audio. Eng. Soc.*, 47(4):218–239, 1999.

[103] D. H. Hwang and J. S. Choi. Real-time binaural sound source localization using sparse coding and SOM. In *Intelligent Robotics and Applications*, pages 582–589. Springer, 2010.

[104] I. Ibrahim, V. Parsa, E. Macpherson, and M. Cheesman. Evaluation of speech intelligibility and sound localization abilities with hearing aids using binaural wireless technology. *Audiol. Res.*, 3(1):1–21, 2013.

[105] A. Ihlefeld and B. G. Shinn-Cunningham. Effect of source spectrum on sound localization in an everyday reverberant room. *J. Acoust. Soc. Am.*, 130(1):324–333, 2011.

[106] International Telecommunications Union (ITU). ITU-T P.800 - Methods for subjective evaluation of transmission quality. *Recommendation ITU-T*, 1996.

[107] International Telecommunications Union (ITU). BS. 1534-1. Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA). *Recommendation ITU-R*, 2001.

[108] H. M. Jackson and B. C. J. Moore. Contribution of temporal fine structure information and fundamental frequency separation to intelligibility in a competing-speaker paradigm. *J. Acoust. Soc. Am.*, 133(4):2421–2430, 2013.

[109] J. Jensen, M. S. Pedersen, M. Farmani, P. Minnaar, and Oticon A/S. Hearing system, 2016. US2016/0112811 A1.

[110] S. Kalluri, K. Fitz, J. Ellison, D. J. Reynolds, and Starkey Hearing Technologies. Spatial enhancement mode for hearing aids, 2016.

[111] B. F. G. Katz and M. Noisternig. A comparative study of interaural time delay estimation methods. *J. Acoust. Soc. Am.*, 135(6):3530–3540, 2014.

[112] T. Kaufmann, J. Sterkens, and J. M. Woodgate. Hearing loops, the preferred assistive listening technology. *J. Audio. Eng. Soc.*, 63(4):298–302, 2015.

[113] G. Keidser, K. Rohrseitz, H. Dillon, V. Hamacher, L. Carter, U. Rass, and El. Convery. The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers. *Int. J. Audiol.*, 45(10):563–579, 2006.

[114] G. Keidser, L. Carter, J. Chalupper, and H. Dillon. Effect of low-frequency gain and venting effects on the benefit derived from directionality and noise reduction in hearing aids. *Int. J. Audiol.*, 46(10):554–568, 2009.

[115] G. Keidser, A. O'Brien, J.-U. Hain, M. McLelland, and I. Yeend. The effect of frequency-dependent microphone directionality on horizontal localization performance in hearing-aid users. *Int. J. Audiol.*, 48(11):789–803, 2009.

[116] G. Keidser, H. Dillon, M. Flax, T. Ching, and S. Brewer. The NAL-NL2 prescription procedure. *Audiol. Res.*, 1(1):88–90, 2011.

[117] F. Keyrouz and K. Diepold. An enhanced binaural 3D sound localization algorithm. In *Proceedings of the International Symposium on Signal Processing and Information Technology (ISSPIT) 2006*, pages 662–665. Institute of Electrical and Electronics Engineers, 2006.

[118] F. Keyrouz and K. Diepold. A rational HRTF interpolation approach for fast synthesis of moving sound. In *Proceedings of the 12th Digital Signal Processing Workshop, 4th Signal Processing Education Workshop (DSP/SPE)*, pages 222–226. Institute of Electrical and Electronics Engineers, 2006.

[119] F. Keyrouz, Y. Naous, and K. Diepold. A new method for binaural 3D localization based on HRTFs. In *Proceedings of the International Conference on Acoustics Speech and*

Signal Processing (ICASSP) 2006, volume 5, pages 341–344. Institute of Electrical and Electronics Engineers, 2006.

[120] A. I. Khuri and M. Conlon. Simultaneous optimization of multiple responses represented by polynomial regression functions. *Technometrics*, 23(4):363–375, 1981.

[121] D. J. Kistler and F. L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Am.*, 91(3):1637–1647, 1991.

[122] C. Knapp and C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.*, 24(4):320–327, 1976.

[123] S. Kobler and U. Rosenhall. Horizontal localization and speech intelligibility with bilateral and unilateral hearing aid amplification. *Int. J. Audiol.*, 41(7):395–400, 2002.

[124] S. Kochkin. MarkeTrak VIII: Consumer satisfaction with hearing aids is slowly increasing. *Hear. J.*, 63(1):19–27, 2010.

[125] S. Kochkin. MarkeTrak VIII Patients report improved quality of life with hearing aid usage. *Hear. J.*, 61(6):25–32, 2011.

[126] B. Kollmeier, T. Brand, R. Huber, and V. Hohmann. Modeling hearing impairment and its effect on auditory quality by auralization. In *Proceedings of the 47th AES Conference*. Audio Engineering Society, 2012.

[127] G. F. Kuhn. Model for the interaural time differences in the azimuthal plane. *J. Acoust. Soc. Am.*, 62(1):157–167, 1977.

[128] A. Kulkarni and H. S. Colburn. Efficient finite impulse response filter models of the head-related transfer function. *J. Acoust. Soc. Am.*, 97(5):3278–3278, 1995.

[129] National Acoustic Laboratories. Hearing loss prevention (protection), 2016. URL http://www.nal.gov.au/hearing-loss-protection_tab_noise-exposure.shtml. 2016-04-21.

[130] M. Latzel. Concepts for binaural processing in hearing aids, 2013. URL http://www.hearingreview.com/2013/03/concepts-for-binaural-processing-in-hearing-aids/. 2016-03-08.

[131] H. T. Lawless and H. Heymann. *Sensory evaluation of food - Principles and practices*. Springer, 1998.

[132] S. Le Bagousse, C. Colomes, and M. Paquier. State of the art on subjective assessment of spatial sound quality. In *Proceedings of the 38th AES Conference*. Audio Engineering Society, 2010.

[133] S. Le Bagousse, M. Paquier, C. Colomes, and S. Moulin. Sound quality evaluation based on attributes-application to binaural contents. In *Proceedings of the 131st AES Convention*. Audio Engineering Society, 2011.

[134] F. Legent, P. Bordure, C. Calais, and O. Malard. *Audiologie pratique: manuel pratique des tests de l'audition*. Masson, 2002.

[135] H. Levitt. Digital hearing aids: a tutorial review. *J. Rehab. Res. Dev.*, 24(4):7–20, 1987.

[136] M. S. Lewis, C. C. Crandell, M. Valente, and J. E. Horn. Speech perception in noise: Directional microphones versus frequency modulation (FM) systems. *J. Am. Acad. Audiol.*, 15(6):426–439, 2004.

[137] D. Li and S. E. Levinson. A linear phase unwrapping method for binaural sound source localization on a robot. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, volume 1, pages 19–23. Institute of Electrical and Electronics Engineers, 2002.

[138] D. Li and S. E. Levinson. A Bayes-rule based hierarchical system for binaural sound source localization. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP) 2003*, volume 5, pages 521–524. Institute of Electrical and Electronics Engineers, 2003.

[139] C. Lim and R. O. Duda. Estimating the azimuth and elevation of a sound source from the output of a cochlear model. In *Proceedings of the Asimolar Conference Signals on Systems and Computers (ASIMOLARSSC) 1994*, volume 1, pages 399–403. Institute of Electrical and Electronics Engineers, 1994.

[140] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *J. Acoust. Soc. Am.*, 106(4):1633–1654, 1999.

[141] J. J. Lopez, E. Aguilera, and P. Gutierrez. A study of the influence of auralization on speech intelligibility and immersion in multi-party teleconferencing systems using binaural audio. In *Proceedings of the Forum Acusticum*. European Acoustics Association, 2014.

[142] C. Lorenzi, S. Gatehouse, and C. Lever. Sound localization in noise in hearing-impaired listeners. *J. Acoust. Soc. Am.*, 105(6):3454–3463, 1999.

[143] T. Lundstedt, E. Seifert, L. Abramo, B. Thelin, A. Nyström, J. Pettersen, and R. Bergman. Experimental design and optimization. *Chemometr. Intell. Lab.*, 42(1):3–40, 1998.

[144] X. Luo, W. J. O'Brien, and C. L. Julien. Comparative evaluation of Received Signal-Strength Index (RSSI) based indoor localization techniques for construction jobsites. *Adv. Gen. Inf.*, 25(2):355–363, 2011.

[145] J. A. MacDonald. A localization algorithm based on head-related transfer functions. *J. Acoust. Soc. Am.*, 123(6):4290–4296, 2008.

[146] J. Mackenzie, J. Huopaniemi, V. Valimaki, and I. Kale. Low-order modeling of head-related transfer functions using balanced model truncation. *IEEE Signal Proc. Let.*, 4(2): 39–41, 1997.

[147] P. Majdak, T. Walder, and B. Laback. Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions. *J. Acoust. Soc. Am.*, 134(3):2148–2159, 2013.

[148] P. Marmaroli, M. Carmona, J. M. Odobez, X. Falourd, and H. Lissek. Observation of vehicle axles through pass-by noise: A strategy of microphone array design. *IEEE Trans. Intell. Transport. Syst.*, 14(4):1654–1664, 2013.

[149] F. N. Martin and J. G. Clark. *Introduction to audiology*. Allyn and Bacon Boston, 2012.

[150] M. C. Martin and I. Summers. *Dictionary of Hearing*. Whurr, 1999.

[151] C. Masterson, S. Adams, G. Kearney, and F. Boland. A method for head related impulse response simplification. In *Proceedings of the 17th European Signal Processing Conference (EUSIPCO)*. European Association for Signal Processing, 2009.

[152] Mathworks. Matlab coder - generate C and C++ code from Matlab code, 2016. URL http://www.mathworks.com/products/matlab-coder/index.html?s_tid=gn_loc_drop.

[153] Mathworks. dfilt.df2t, 2016. URL http://ch.mathworks.com/help/dsp/ref/dfilt.df2t.html.

[154] Mathworks. equiripple, 2016. URL http://ch.mathworks.com/help/signal/ref/equiripple.html.

[155] Mathworks. fir1, 2016. URL http://ch.mathworks.com/help/signal/ref/fir1.html.

[156] Mathworks. fir2, 2016. URL http://ch.mathworks.com/help/signal/ref/fir2.html.

[157] Mathworks. firls, 2016. URL http://ch.mathworks.com/help/signal/ref/firls.html.

[158] Mathworks. iirls, 2016. URL http://ch.mathworks.com/help/dsp/ref/iirls.html.

[159] Mathworks. prony, 2016. URL http://ch.mathworks.com/help/signal/ref/prony.html.

[160] Mathworks. yulewalk, 2016. URL http://ch.mathworks.com/help/signal/ref/yulewalk.html.

[161] Hearing Health & Technology Matters. Wireless systems for hearing aids, 2013. URL http://hearinghealthmatters.org/waynesworld/2013/2566/. 2016-03-08.

[162] Hearing Health & Technology Matters. Wireless systems for hearing aids: part II, 2013. URL http://hearinghealthmatters.org/waynesworld/2013/wireless-systems-for-hearing-aids-part-ii/. 2016-03-08.

[163] Hearing Health & Technology Matters. If hearing aids are not enough, FM systems may be the solution, 2015. URL http://hearinghealthmatters.org/hearinprivatepractice/2015/if-hearing-aids-are-not-enough-fm-systems-may-be-the-solution/. 2016-03-08.

[164] A. McCormack and H. Fortnum. Why do people fitted with hearing aids not wear them? *Int. J. Audiol.*, 52(5):360–368, 2013.

[165] R. W. McCreery. The effects of frequency-lowering on speech understanding in children. *Hear. J.*, 65(7):14, 2012.

[166] R. W. McCreery, J. Alexander, M. A. Brennan, B. Hoover, J. Kopun, and P. G. Stelmachowicz. The influence of audibility on speech recognition with nonlinear frequency compression for children and adults with hearing loss. *Ear and Hear.*, 35(4):440–447, 2014.

[167] C. Mendonça, G. Campos, P. Dias, J. Vieira, J. P. Ferreira, and J. A. Santos. On the improvement of localization accuracy with non-individualized HRTF-based sounds. *J. Audio. Eng. Soc.*, 60(10):821–830, 2012.

[168] P. Minnaar, J. Plogsties, and F. Christensen. Directional resolution of head-related transfer functions required in binaural synthesis. *J. Acoust. Soc. Am.*, 53(10):919–929, 2005.

[169] J. J. M. Monaghan, K. Krumbholz, and B. U. Seeber. Factors affecting the use of envelope interaural time differences in reverberation. *J. Acoust. Soc. Am.*, 133(4):2288–2300, 2013.

[170] B. C. J. Moore. *Hearing*. Academic Press, 1995.

[171] B. C. J. Moore. *Cochlear hearing loss: physiological, psychological and technical issues*. John Wiley & Sons, 2007.

[172] B. C. J. Moore. *An introduction to the psychology of hearing*. Emerald, 2012.

[173] M. F. Mueller, A. Kegel, S. M. Schimmel, N. Dillier, and M. Hofbauer. Localization of virtual sound sources with bilateral hearing aids in realistic acoustical scenes. *J. Acoust. Soc. Am.*, 131(6):4732–4742, 2012.

[174] C. Neti, E. D. Young, and M. H. Schneider. Neural network models of sound localization based on directional filtering by the pinna. *J. Acoust. Soc. Am.*, 92(6):3140–3156, 1992.

[175] L. B. Nielsen. Subjective evaluation of sound quality for normal-hearing and hearing-impaired Listeners. Technical Report #51, The Acoustics Laboratory - Technical University of Denmark (TUD), 1992.

[176] M. Nilsson, S. D. Soli, and J. A. Sullivan. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.*, 95(2):1085–1099, 1994.

[177] J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *J. Acoust. Soc. Am.*, 119(1):463–479, 2006.

[178] W. Noble and D. Byrne. A comparison of different binaural hearing aid systems for sound localization in the horizontal and vertical planes. *Br. J. Audiol.*, 24(5):335–346, 1990.

[179] W. Noble, D. Byrne, and B. Lepage. Effects on sound localization of configuration and type of hearing impairment. *J. Acoust. Soc. Am.*, 95(2):992–1005, 1994.

[180] W. Noble, D. Byrne, and K. Ter-Horst. Auditory localization, detection of spatial separateness, and speech hearing in noise by hearing impaired listeners. *J. Acoust. Soc. Am.*, 102(4):2343–2352, 1997.

[181] Y. Oesch. D1.1 Specifications of PHONAK BHA(L&S) prototype - Restricted. KTI deliverable, Phonak Communications AG, 2012.

[182] Y. Oesch. D4.4 Verification of the portable mock-up, Body Worn Unit V2 - Restricted. KTI deliverable, Phonak Communications AG, 2014.

[183] Y. Oesch, C. Richard, T. Jost, M. Secall, C. Schmid, R. Platz, E. Dijkstra, and Phonak AG. Hearing assistance system and method, 2011. WO 2011/015675 A2.

[184] B. Ohl, S. Laugesen, J. Buchholz, and T. Dau. Externalization versus internalization of sound in normal-hearing and hearing-impaired listeners. In *Proceedings DEGA*. Deutschen Gesellschaft für Akustik, 2010.

[185] World Health Organization. Prevention of blindness and deafness, 2016. URL http://www.who.int/pbd/deafness/hearing_impairment_grades/en/. 2016-03-08.

[186] T. Petsatodis, F. Talantzis, C. Boukis, Z. Tan, and R. Prasad. Exploring super-gaussianity toward robust information-theoretical time delay estimation. *J. Acoust. Soc. Am.*, 133: 1515–1524, 2013.

[187] E. M. Picou, E. Aspell, and T. A. Ricketts. Potential benefits and limitations of three types of directional processing in hearing aids. *Ear and Hear.*, 35(3):339–352, 2014.

[188] H. J. Platte and P. Laws. Die Vorneortung bei der kopfbezogenen Stereophonie (Frontal localization in head-related stereophony). *Rad. Mentor Electron.*, 42:97–10, 1976.

[189] G. Plenge. On the differences between localization and lateralization. *J. Acoust. Soc. Am.*, 56(3):944–951, 1974.

[190] R. Plomp. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J. Acoust. Soc. Am.*, 63(2):533–549, 1978.

[191] R. Plomp and A. M. Mimpen. Speech-reception threshold for sentences as a function of age and noise level. *J. Acoust. Soc. Am.*, 66(5):1333–1342, 1979.

[192] I. Pollack and J. M. Pickett. Stereophonic listening and speech intelligibility against voice babble. *J. Acoust. Soc. Am.*, 30(2):131–133, 1958.

[193] T. A. Powers and P. Burton. Wireless technology designed to provide true binaural amplification. *Hear. J.*, 58(1):25–34, 2005.

[194] P. Prandoni and M. Vetterli. *Signal processing for communications*. CRC Press, 2008.

[195] M. Queiroz and G.H.M. De Sousa. *Structured IIR models for HRTF Interpolation*. University of Michigan Library Repository, 2010.

[196] A. Raake and B. F. G. Katz. SUS-based method for speech reception threshold measurement in French. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. Association of European Language Ressources, 2006.

[197] B. Rakerd and W. M. Hartmann. Localization of sound in rooms. V. Binaural coherence and human sensitivity to interaural time differences in noise. *J. Acoust. Soc. Am.*, 128(5): 3052–3063, 2010.

[198] J. Ramirez, J. M. Górriz, and J. C. Segura. Voice activity detection. fundamentals and speech recognition system robustness. In *Robust speech recognition and understanding*, pages 1–22. I-Tech Education and Publishing, 2007.

[199] M. Raspaud, H. Viste, and G. Evangelista. Binaural source localization by joint estimation of ILD and ITD. *IEEE Trans. Audio Speech Lang. Process.*, 18(1):68–77, 2010.

[200] C. Renard and Mikael Menard. Les évolutions techniques des aides auditives. *Acoustique & techniques*, (56):24–35, 2009.

[201] Hearing Review. National health survey finds 1 in 6 adults has trouble hearing, 2015. URL http://www.hearingreview.com/2015/09/national-health-survey-finds-1-6-adults-trouble-hearing/. 2016-07-05.

[202] Hearing Review. Study reveals how hearing technology helps children, 2015. URL http://www.hearingreview.com/2016/03/study-reveals-how-hearing-technology-helps-children/. 2016-07-05.

[203] Hearing Review. Silently suffering with hearing loss negatively affects quality of life, 2015. URL http://www.hearingreview.com/2015/08/silently-suffering-hearing-loss-negatively-affects-quality-life/. 2016-07-05.

[204] Hearing Review. US hearing aid unit sales grow by 10% in Q2 of 2016, 2015. URL http://www.hearingreview.com/2016/07/us-hearing-aid-unit-sales-grow-10-q2-2016/.

[205] Hearing Review. SamMobile hints at Samsung's launch into hearing aid market, 2015. URL http://www.hearingreview.com/2016/01/sammobile-hints-samsungs-launch-hearing-aid-market/. 2016-07-07.

[206] C. Richard and Y. Oesch. Mercury spatial hearing investigations - Restricted technical note R&D. Internal report, Phonak Communications AG, 2010.

[207] L. Rohr. *Evaluation of audio source separation in the context of 3D audio*. PhD thesis, EPFL, 2015.

[208] F. Rumsey. Music-induced hearing disorders. *J. Audio. Eng. Soc.*, 60(11):965–968, 2012.

[209] P. R. Runkle, M. A. Blommer, and G. H. Wakefield. A comparison of head related transfer function interpolation methods. In *Proceedings of the workshop on Applications of Signal Processing to Audio and Acoustics (ASSP)*, pages 88–91. Institute of Electrical and Electronics Engineers, 1995.

[210] M. Sahidullah and G. Saha. Comparison of speech activity detection techniques for speaker recognition, 2012.

[211] J. Sandvad and D. Hammershøi. What is the most efficient way of representing HTF filters? In *Proceedings of the Nordic Signal Processing Symposium (NORSIG)*, pages 174–178. Institute of Electrical and Electronics Engineers, 1994.

[212] C. Schauer, T. Zahn, P. Paschke, and H. Gross. Binaural sound localization in an artificial neural network. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP) 2000*, volume 2, pages 865–868. Institute of Electrical and Electronics Engineers, 2000.

[213] S. Scollie, R. Seewald, L. Cornelisse, S. Moodie, M. Bagatto, D. Laurnagaray, S. Beaulac, and J. Pumford. The desired sensation level multistage input/output algorithm. *Trends Amplif.*, 9(4):159–197, 2005.

[214] L. R. Shanock, B. E. Baran, W. A. Gentry, S. C. Pattison, and E. D. Heggestad. Polynomial regression with response surface analysis: A powerful approach for examining moderation and overcoming limitations of difference scores. *J. of Bus. Psychol.*, 25(4):543–554, 2010.

[215] A. Simpson, H. J. McDermott, and R. C. Dowell. Benefits of audibility for listeners with severe high-frequency hearing loss. *Hear. R.*, 210(1):42–52, 2005.

[216] C. Smits and J. M. Festen. The interpretation of speech reception threshold data in normal-hearing and hearing-impaired listeners: Steady-state noise. *J. Acoust. Soc. Am.*, 130(5):2987–2998, 2011.

[217] C. Smits and J. M Festen. The interpretation of speech reception threshold data in normal-hearing and hearing-impaired listeners: II. Fluctuating noise. *J. Acoust. Soc. Am.*, 133(5):3004–3015, 2013.

[218] R. Sockalingam, M. Holmberg, K. Eneroth, and M. Shulte. Binaural hearing aid communication shown to improve sound quality and localization. *Hear. J.*, 62(10):46–47, 2009.

[219] M. Srbinovska, C. Gavrovski, and V. Dimcev. Localization estimation system using measurement of RSSI based on ZigBee standard. In *Proceedings of ELECTRONICS 2008*. Higher Education Forum, 2008.

[220] M. A. Stone, K. Anton, and B. C. J. Moore. Use of high-rate envelope speech cues and their perceptually relevant dynamic range for the hearing impaired. *J. Acoust. Soc. Am.*, 132(2):1141–1151, 2012.

[221] K. E. Strom. Hearing aid sales increase by 4.8% in 2014; RICs continue market domination, 2015. URL http://www.hearingreview.com/2015/01/hearing-aid-sales-increase-4-8-2014-rics-continue-market-domination/. 2016-07-05.

[222] Confédération Suisse. Article constitutionnel 118b, 2007.

[223] Confédération Suisse. Loi relative à la recherche sur l'être humain (LRH), 2009.

[224] Confédération Suisse. Ordonnance sur les essais cliniques (OClin), 2013.

[225] Swissethics. http://swissethics.ch/, 2016.

[226] D. S. Talagala, W. Zhang, T. D. Abhayapala, and A. Kamineni. Binaural sound source localization using the frequency diversity of the head-related transfer function. *J. Acoust. Soc. Am.*, 135(3):1207–1217, 2014.

[227] B. Taylor. International speech test signal (ISTS). URL http://www.ehima.com/. 2015-06-01.

[228] B. Taylor. Research firm analyzes market share, retail activity, and prospects of major hearing aid manufacturers, 2013. URL http://hearinghealthmatters.org/hearingnewswatch/2013/research-firm-analyzes-market-share-retail-stores-prospects-of-major-hearing-aid-makers/. 2016-07-05.

[229] L. Thibodeau. Benefits of adaptive FM systems on speech recognition in noise for listeners who use hearing aids. *Am. J. Audiol.*, 19(1):36–45, 2010.

[230] W. R. Thurlow and P. S. Runge. Effect of induced head movements on localization of direction of sounds. *J. Acoust. Soc. Am.*, 42(2):480–488, 1967.

[231] B. Timmer. It's sync or stream! The differences between wireless hearing aid features, 2013. URL http://www.hearingreview.com/2013/05/it-s-sync-or-stream-the-differences-between-wireless-hearing-aid-features/. 2016-03-08.

[232] T. Van den Bogaert, T. J. Klasen, M. Moonen, L. Van Deun, and J. Wouters. Horizontal localization with bilateral hearing aids: without is better than with. *J. Acoust. Soc. Am.*, 119(1):515–526, 2006.

[233] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen. The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids. *J. Acoust. Soc. Am.*, 124(1):484–497, 2008.

[234] T. Van den Bogaert, E. Carette, and J. Wouters. Sound localization with and without hearing aids. In *Proceedings ICA*, pages 1314–1317. International Commission for Acoustics, 2009.

[235] T. Van den Bogaert, E. Carette, and J. Wouters. Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna. *Int. J. Audiol.*, 50(3):164–176, 2011.

[236] S. Van Hemel and R. A. Dobie. *Hearing loss: determining eligibility for social security benefits*. National Academies Press, 2004.

[237] J. Vieira and L. Almeida. A sound localizer robust to reverberation. In *Proceedings of the 115th AES Convention*. Audio Engineering Society, 2003.

[238] E. M. Von Hornbostel and M. Wertheimer. Über die Wahrnehmung der Schallrichtung. *Sitzungsberichte der preussischen Akademie der Wissenschaften*, 388:396, 1920.

[239] X. Wan and J. Liang. Robust and low complexity localization algorithm based on head-related impulse responses and interaural time difference. *J. Acoust. Soc. Am.*, 133(1): EL40–EL46, 2013.

[240] L. Wang, F. Yin, and Z. Chen. Head-related transfer function interpolation through multivariate polynomial fitting of principal component weights. *Acoust. Sci. Technol.*, 30(6):395–403, 2009.

[241] R. M. Warren. *Auditory perception: an analysis and synthesis*. Cambrige, 2008.

[242] G. Waters. Sound quality assessment material recordings for subjective tests (SQAM), 1988.

[243] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.*, 94(1):111–123, 1993.

[244] W. M. Whitmer, B. U. Seeber, and M. A. Akeroyd. The perception of apparent auditory source width in hearing-impaired adults. *J. Acoust. Soc. Am.*, 135(6):3548–3559, 2014.

[245] I. M. Wiggins and B. U. Seeber. Effects of dynamic-range compression on the spatial attributes of sounds in normal-hearing listeners. *Ear and Hear.*, 33(3):399–410, 2012.

[246] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. I: Stimulus synthesis. *J. Acoust. Soc. Am.*, 85(2):858–867, 1989.

[247] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. II: Psychophysical validation. *J. Acoust. Soc. Am.*, 85(2):868–878, 1989.

[248] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Körner. A probabilistic model for binaural sound localization. *IEEE Trans. Syst. Man. Cybern.*, 36(5):982–994, 2006.

[249] R. S. Woodworth and H. Schlosberg. *Experimental psychology*. Oxford and IBH Publishing, 1954.

[250] W. A. Yost. *Fundamentals of hearing: an introduction*. Emerald, 2006.

[251] N. Yousefian, P. C. Loizou, and J. H. L. Hansen. A coherence-based noise reduction algorithm for binaural hearing aids. *Speech Commun.*, 58:101–110, 2014.

[252] C. Zhou, R. Hu, W. Tu, X. Wang, and L. Gao. Binaural moving sound source localization by joint estimation of ITD and ILD. In *Proceedings of the 130th AES Convention*. Audio Engineering Society, 2011.

[253] T. S. Zurbrügg, A. Stirnemann, M. Kuster, and H. Lissek. Investigations on the physical factors influencing the occlusion effect in ear canal occlusion effect caused by hearing aids. *Acta Acust. united Ac.*, 2013.

[254] E. Zwicker, R. Feldtkeller, and C. Sorin. *Psychoacoustique: L'oreille, récepteur d'information*. Masson, 1981.

Gilles Courtois

PhD Candidate in Hearing Technologies

Personal Data

Place and Date of Birth:	Thann, France | 16 November 1989
Address:	Chemin de l'Orme 2, 1054 Morrens, Switzerland
Nationality:	French
Marital Status:	With partner, No children
Phone:	+41 (0)78 639 52 62
email:	gilles.courtois@epfl.ch
Expertise:	Audio signal processing, Audiology, Psychoacoustics

Work Experience

From Sept 2012 — Research assistant at EPFL (Ecole Polytechnique Fédérale de Lausanne), Switzerland

PhD candidate in hearing technologies, Laboratory of Signal Processing (Prof. Pierre Vandergheynst)

Research and development of an algorithm for remote microphone systems for hearing aids. Collaboration with Phonak Communications AG, Murten, Switzerland:

- Development of an algorithm for localizing a speaker and spatializing the voice through the hearing aids of a listener.
- Porting of the algorithm on a microprocessor Atem ARM.
- Psychoacoustic test guidance with 40 normal-hearing subjects.

Direction of a clinical trial as a Principal Investigator. Collaboration with Sonova AG (Phonak AG):

- Evaluation of a speech spatialization method for hearing aids. Cohort of 40 normal-hearing and hearing-impaired patients.
- Complete protocol writing and submission of the research project to an ethics committee.
- Guidance and management of the experiment. Collaboration with an audiologist. Recruitment and interactions with the patients.
- Analysis of the results and dissemination.

Supervision of a Tunisian engineering team for a European Project. Collaboration with the Fiat Research Center (Italy) and Active Audio Corporation (France):

- Design of digital filters to improve intelligibility of presbycusic subjects.
- Development of a microphone array to improve car driving for presbycusis subjects.
- Psychoacoustics test direction with 20 normal-hearing and hearing-impaired subjects. Collaboration with an ENT doctor.

Feb-Jul 2012 — Internship in audio signal processing, EPFL, Switzerland

Master thesis, AudioVisual Communications Lab (Prof. Martin Vetterli)

Work on a psychoacoustic model to simulate the perception and the localization ability of the human auditory system within the context of a stereophonic rendering.

Feb-Jul 2011 — Engineering internship in electroacoustics, Focal JMlab Corporation, Saint-Etienne, France

Research and development in the car audio and HiFi teams

Study for the use of a KEMAR dummy head to improve and stabilize the perceived sound stage of the driver. Psychoacoustic evaluations. Set up of a procedure for fast measurements of loudspeaker directivity.

Education

Oct 2016 **Doctor of Philosophy (PhD)** - EPFL, Switzerland (Expected)
Thesis: "Development of binaural hearing aids towards the rendering of auditory spatial perception" | Advisor: Dr Hervé Lissek

- International conferences (AES conventions, EUSIPCO, Interspeech).
- Advanced signal processing and financial management courses.
- Supervision of bachelor students and interns.
- Certificate of Good Clinical Practice (GCP).

July 2012 **Master of Science (M.Sc.)** and **Master of Engineering (M.Eng.)** - INSA (Institut National des Sciences Appliquées), Lyon, France - Passed with honors
Thesis: "A model to predict phantom source localization in different listening positions" | Advisor: Dr Andreas Walther

- Main courses: Signal processing, analog electronics, embedded systems, acoustics, C-programming language, project management.
- 1-year exchange study (Erasmus) at EPFL for a specialization in signal processing for audio and acoustics.
- Semester projects in the field of analog electronics for audio applications.

July 2007 Baccalauréat in Science (French high-school degree) - Passed with honors

Languages

French: Mothertongue
English: Fluent (TOEIC score: 910, FCE of Cambridge University: Grade C)
German: Moderate
Spanish: Basis

Computer Skills

Programming:	MATLAB, Simulink, C-code, MATLAB Coder toolbox (Floating-point to fixed-point conversion and automatic C-code generation), Python (Basis)
Audiology:	Phonak Target, Phonak iPFG
Multimedia:	Adobe Audition, Adobe Illustrator, Microsoft Visio, Final Cut Pro X, Cubase Pro
Statistics:	SPSS
Operating Systems :	Windows, Mac OS

Interests and Activities

Audiology, Technology
Music, Reading
Hiking, Badminton

Patent

Jan 2015 Hearing assistance system, International, WO/2016/116160, G Courtois, P. Marmaroli, H. Lissek, Y. Oesch, W. Balande.

Publications

Sept 2016	Dynamic range limiting of HRTFs: principle and objective evaluation, G. Courtois, P. Marmaroli, L. Rohr, H. Lissek, Y. Oesch, W. Balande, J. Audio Eng. Soc., 64(10), 2016.
Sept 2015	Development and assessment of a localization algorithm implemented in binaural hearing aids, G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, W. Balande, 23rd European Signal Processing Conference (EUSIPCO), Nice, France.
Sept 2015	Intelligibility enhancement of vocal announcements for public address systems: a design for all through a presbycusis pre-compensation Filter, A. Ben Jemaa, N. Mechergui, G. Courtois, A. Mudry, S. Djaziri Larbi, M. Turki, H. Lissek, M. Jaidane, Interspeech, Dresden, Germany.
May 2015	Binaural hearing aids with wireless microphone systems including speaker localization and spatialization, G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, W. Balande, AES 138th Convention, Warsaw, Poland.
May 2014	Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions, G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, W. Balande, AES 136th Convention, Berlin, Germany.
April 2014	Implémentation d'un algorithme de localisation binaurale sur audioprothèses : Contraintes et perspectives, G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, W. Balande, CFA, Poitiers, France.

References

Dr Hervé Lissek
Head of the acoustic group
EPFL
Phone: +41 21 693 46 30

Mr Yves Oesch
Project manager
Phonak Communications AG
Phone: +41 26 672 96 72

Dr Xavier Gigandet
DSP engineer
Phonak Communications AG
Phone: +41 26 672 96 72

Mr Philippe Estoppey
Audiologist
Acoustique Riponne
Phone: +41 21 320 61 36